

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.

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 $s = 1 + \lfloor \log_2 n \rfloor$ або $s = \lfloor 10 \log_{10}(n) \rfloor$;
3. $x_i^* = \frac{a_i + b_i}{2}$ – центральна точка.

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

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

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 під-інтервалу якому належить x .

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 мають 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”.

Програмні датчики

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

Лінійна змішана формула

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

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

Мультиплікативний конгруентний метод

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

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

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

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

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

Теорема 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

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

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

Ускладнення лінійної змішаної формули

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

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

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 $\lfloor 1 + mx \rfloor$.

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 $U([0, 1])$.

Моделювання нормального розподілу з параметрами m та σ^2

Теорема 3. 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 $a + (b - a)x$.

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

$$\begin{aligned} 1 - F(y_{i(Q)}) &= P\{\xi \geq y_{i(Q)}\} > Q/100, \\ 1 - F(y_{i(Q)+1}) &= P\{\xi \geq y_{i(Q)+1}\} \leq Q/100. \end{aligned}$$

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 $(u_{3/4} - u_{1/4})$.

Ймовірнісне відхилення d_ξ визначається як половина інтерквартильної широти, тобто $d_\xi = \frac{u_{3/4} - u_{1/4}}{2}$.

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 $(u_{5/6} - u_{1/6})$.

Характеристики положення центра значень змінної

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

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

Середнє геометричне G_ξ визначається для випадкових величин, які з ймовірністю 1 додатні. Згідно визначення $G_\xi = \exp\{M(\ln(\xi))\}$. Оцінка величини має 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 так: $H_\xi = 1/M(1/\xi)$. Емпіричне значення середнього гармонічного має вигляд $\hat{H}_\xi = n / \sum_{i=1}^n \frac{1}{x_i}$.

Мода x_{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_{med} – це квантиль рівня $1/2$, її оцінка $\hat{x}_{\text{med}}(n)$ обчислюється на основі емпіричної функції розподілу.

Характеристики розсіювання значень змінної

Маємо вибірку об'єму n спостережень x_1, x_2, \dots, 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 $s^2(n) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}(n))^2$. Деколи більш корисно представляти $s^2(n) = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2(n) \right)$.

Стандартне (середнє квадратичне) відхилення σ є коренем з дисперсії $\sigma = \sqrt{D\xi}$.

$$s(n) = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2(n) \right)}.$$

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 $\hat{e}_\xi = s(n)/\sqrt{n}$.

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

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

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

Розмах (широта) вибірки x_1, x_2, \dots, x_n спостережень над ξ визначається таким чином: $\widehat{R}_\xi(n) = x_{\max}(n) - x_{\min}(n)$, 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 $(\bar{x}(n) - 3s(n), \bar{x}(n) + 3s(n))$.

Аналіз скошеності та гостроверхості розподілу

Маємо вибірку об'єму n спостережень x_1, x_2, \dots, 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** – характеристики скошеності розподілу – лежить третій центральний момент

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

Вибіркове значення:

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

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

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

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

Характеристики випадкових векторів

Маємо q -мірний випадковий вектор ξ .

Отримано n спостережень $x_1, x_2, \dots, x_n \in \mathbb{R}^q$, $i = \overline{1, n}$.

Характеристики положення центра значень

Математичне сподівання представляє собою вектор, а формула обчислень лишається без змін: $M\xi$, вибіркове значення $\bar{x}(n) = \frac{1}{n} \sum_{i=1}^n x_i$.

Мода x_{mod} у випадку вектора визначається як точка максимуму щільності ξ , для дискретного випадку це значення ξ яке набувається з максимальною ймовірністю.

Характеристики розсіювання значень

Коваріаційна матриця визначається як $\Sigma = M(\xi - M\xi)(\xi - M\xi)^T$. Емпіричний варіант:

$$\widehat{\Sigma}(n) = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}(n))(x_k - \bar{x}(n))^T$$

Узагальнена дисперсія визначається як визначник коваріаційної матриці: $\det \Sigma$, емпіричний вигляд $\det(\widehat{\Sigma}(n))$.

Зауваження! Слід (trace) для квадратної матриці A рівний $\text{tr}(A) = \sum_{i=1}^n A_{i,i}$.

Слід коваріаційної матриці підраховується як $\text{tr}(\Sigma)$, емпіричний вигляд $\text{tr}(\widehat{\Sigma}(n))$.

Перевірка стохастичності вибірки

Для перевірки стохастичності вибірки існують такі критерії:

- критерій серій на базі медіани вибірки;
- критерій зростаючих та спадаючих серій;
- критерій квадратів послідовних різниць (критерій Аббе).

Нехай x_1, x_2, \dots, x_n – вибірка спостережень, яка досліджується. Будемо перевіряти гіпотезу про те, що ця вибірка є випадковою з рівнем значущості α ($0 < \alpha < 1$).

Критерій серій на базі медіани вибірки

1. Визначається оцінка медіани \hat{x}_{med} .
2. Під кожним членом вибірки ставиться плюс якщо він строго більший за медіану і мінус навпаки. Виміри які дорівнюють медіані не враховуються.
3. Для послідовності плюсів та мінусів обчислюється загальна кількість серій $\nu(n)$ та кількість членів у найдовшій серії $\tau(n)$. Серією називається під-послідовність однакових знаків.

Зауваження! Вибірка матиме стохастичну природу якщо довжина найдовшої серії $\tau(n)$ на занадто довга, а загальна кількість серій $\nu(n)$ не занадто мала.

Гіпотеза:

$$\begin{cases} \nu(n) > \nu_\beta(n), \\ \tau(n) < \tau_{1-\beta}(n), \end{cases}$$

де $\nu_\beta(n)$, $\tau_\beta(n)$ – квантилі рівня β статистик $\nu(n)$ та $\tau(n)$. При фіксованому β рівень значимості α буде лежати у межах $\beta < \alpha < 2\beta - \beta^2$. Якщо порушується один критерій гіпотеза відхиляється.

Критерій зростаючих та спадаючих серій

Критерій чутливий не тільки до монотонних даних, але й циклічних.

1. Замінити підряд розташовані значення одним представником.
2. Під членом вибірки ставиться плюс або мінус в залежності від того, чи наступний член строго більший або строго менший за даний відповідно.
3. Обчислюємо аналогічні статистики $\nu(n)$ та $\tau(n)$ як і в попередньому критерії.

Гіпотеза:

$$\begin{cases} \nu(n) > \nu_\beta(n), \\ \tau(n) < \tau_{1-\beta}(n), \end{cases}$$

Рівень значущості α лежить в тих же межах $\beta < \alpha < 2\beta - \beta^2$.

Критерій квадратів послідовних різниць (критерій Аббе)

Використовується при роботі з нормальними вибірками. У ролі альтернативи при перевірці нашої гіпотези тут може виступати наявність систематичного зміщення у вибірці.

Порахуємо наступну статистику:

$$\gamma(n) = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{2 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2(n) \right)}$$

Гіпотеза приймається якщо $\gamma(n) > \gamma_\alpha(n)$, де для $n \leq 60$ існують табличні значення квантиля, або ж для $n > 60$:

$$\gamma_\alpha(n) = 1 + \frac{u_\alpha}{\sqrt{n + (1 + u_\alpha^2)/2}},$$

де u_α – квантиль рівня α для $N(0, 1)$.

Рангові критерії однорідності

Розглянемо випадкові величини $\xi_1, \xi_2, \dots, \xi_k$ з функціями розподілу $F_1(x), F_2(x), \dots, F_k(x)$. На їх основі сформуємо об'єднану вибірку $\nu_1, \nu_2, \dots, \nu_n$, а для кожної змінної ξ_i отримаємо незалежні спостереження $x_1^{(i)}, x_2^{(i)}, \dots, x_{n_i}^{(i)}$, $i = \overline{1, k}$. Тоді сформована вибірка буде об'ємом $n = \sum_{i=1}^k n_i$. Для спрощення

вважаємо, що всі виміри ν_i , $i = \overline{1, n}$ різні. Розташувавши ці значення у порядку зростання, отримаємо варіаційний ряд $\nu_{(1)}, \nu_{(2)}, \dots, \nu_{(n)}$. Члени варіаційного ряду називають порядковими статистиками.

Рангом спостереження ν_i ($i = \overline{1, n}$) називається його порядковий номер у побудованому варіаційному ряді, позначається $R_{i,n}$ – ранг спостереження ν_i ($i = \overline{1, n}$).

Статистика для лінійного рангового критерію:

$$K_i = \sum_{j=N_i-n_i+1}^{N_i} \varphi(R_{j,n}), \quad N_i = \sum_{j=1}^i n_j, \quad i = \overline{1, k}$$

K_i – статистика по спостереження над ξ_i , $\varphi(R_{i,n})$ – мітка.

Потрібно перевірити гіпотезу $H_0 : F_1(x) = F_2(x) = \dots = F_k(x)$, $\forall x$ з рівнем значимості α ($0 < \alpha < 1$). Хочемо переконатись зо всі випадкові величини однаково розподілені.

Випадок двох вибірок

Гіпотеза $H_0 : F_1(x) = F_2(x)$, $\forall x$ з рівнем значимості α ($0 < \alpha < 1$). Альтернативні гіпотези:

$$H_{11} : F_1(x) = F_2(x - \Delta), \forall x, (\Delta \neq 0)$$

$$H_{12} : F_1(x) = F_2(x - \Delta), \forall x, (\Delta > 0)$$

$$H_{13} : F_1(x) = F_2(x - \Delta), \forall x, (\Delta < 0)$$

Всі критерії розглядаються над першою змінною ξ_1 .

Додаток. Характеристики порядкових статистик

Нехай ξ – випадкова величина, що є нормально розподіленою з параметрами 0 та 1 з функцією розподілу $\Phi(x)$ та функцією щільності $p(x)$. По вибірці x_1, x_2, \dots, x_n незалежних спостережень над ξ побудуємо варіаційний ряд $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.

Для обчислення математичного сподівання m -ої порядкової статистики $x_{(m)}$ можна застосувати наступну формулу:

$$Mx_{(m)} = \Phi^{-1}(\alpha_m) - \frac{\beta_m}{2} \frac{p'(\alpha_m)}{p^2(\alpha_m)} + \frac{\gamma_m}{6} \frac{2(p'(\alpha_m))^2 - p''(\alpha_m)}{(p'(\alpha_m))^3} + O\left(\frac{m}{n^4}\right),$$

$$\text{де } \alpha_m = \frac{m}{n+1}, \beta_m = \frac{m(n-m+1)}{(n+1)^2(n+2)}, \gamma_m = \frac{2m(n-2m+1)(n-m+1)}{(n+1)^3(n+2)(n+3)}.$$

Або ж більш грубе наближення $Mx_m \approx \Phi^{-1}(\alpha_m)$.

Апроксимуємо тепер $\Phi^{-1}(\alpha)$, $\alpha \in (0, 1)$. Якщо $\alpha \in (0, 0.5)$, то $\Phi^{-1}(\alpha) = -\Phi^{-1}(1 - \alpha)$. Якщо ж $\alpha \in [0.5, 1)$, то

$$\Phi^{-1}(\alpha) = \tau - \frac{a_0 + a_1\tau + a_2\tau^2}{1 + b_1\tau + b_2\tau^2 + b_3\tau^3} + \varepsilon, \quad |\varepsilon| < 4.5 \cdot 10^{-4},$$

$$\tau = \sqrt{-2 \ln(1 - \alpha)}, \quad a_0 = 2.515517, \quad a_1 = 0.802853, \quad a_2 = 0.010328, \quad b_1 = 1.432788, \quad b_2 = 0.189269, \quad b_3 = 0.001308.$$

Критерій нормальних міток (Фішера)

$C = \sum_{i=1}^{n_1} M(R_{i,n}, n)$, де $M(m, n)$ – математичне сподівання m -ої порядкової статистики вибірки довжини $n = n_1 + n_2$ нормально розподіленої величини з параметрами 0 та 1.

Статистика C має наступні характеристики при справедливості нульової гіпотези:

$$MC = 0, DC = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^n (M(i, n))^2$$

Критерій Ван дер Вардена

Статистика критерію має вигляд $V = \sum_{i=1}^{n_1} \Phi^{-1} \left(\frac{R_{i,n}}{n+1} \right)$, де $\Phi^{-1}(x)$ – функція обернена до функції розподілу з параметрами 0 та 1, причому коли справедлива нульова гіпотеза, то

$$MV = 0, DV = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^n \left(\Phi^{-1} \left(\frac{i}{n+1} \right) \right)^2$$

Критерій Вілкоксона

Статистика критерію має вигляд $S = \sum_{i=1}^{n_1} R_{i,n}$, причому коли справедлива нульова гіпотеза, то

$$MS = \frac{n_1(n+1)}{2}, \quad DS = \frac{n_1 n_2 n}{12}$$

Процедура використання статистик C , V , S для перевірки гіпотези H_0 однакова: позначимо через U деяку статистику (C , V , або S), $\bar{U} = \frac{U - MU}{\sqrt{DU}}$. В залежності від альтернативної гіпотези H_0 приймається якщо:

- $|\bar{U}| < u_{1-\alpha/2}$, якщо альтернатива H_{11} ,
- $\bar{U} < u_{1-\alpha}$, якщо альтернатива H_{12} ,
- $\bar{U} > u_{\alpha}$, якщо альтернатива H_{13} ,

де u_{α} – квантиль рівня α для нормального розподілу з параметрами 0 та 1.

По мірі спадання потужності критерії розташовуються так: критерій нормальних міток Фішера, критерій Ван дер Вардена, критерій Вілкоксона.

Загальний випадок

Гіпотеза $H_0: F_1(x) = F_2(x) = \dots = F_k(x), \forall x$ з рівнем значимості α ($0 < \alpha < 1$).

1. Будуємо об'єднану вибірку $\nu_1, \nu_2, \dots, \nu_n$ об'єму $n = \sum_{i=1}^k n_i$, а потім відповідний варіаційний ряд $\nu_{(1)}, \nu_{(2)}, \dots, \nu_{(n)}$.
2. Для кожної ξ_i підрахуємо статистику $K_i = \sum_{j=N_i-n_i+1}^{N_i} \psi(R_{j,n})$.

Для неї підійде будь-яка статистика з попередніх критеріїв:

$$C_i = \sum_{j=N_i-n_i+1}^{N_i} M(R_{j,n}, n), \quad V_i = \sum_{j=N_i-n_i+1}^{N_i} \Phi^{-1} \left(\frac{R_{j,n}}{n+1} \right), \quad s_i = \sum_{j=N_i-n_i+1}^{N_i} R_{j,n}.$$

3. Далі знаходимо їх стандартизовані значення $\bar{K}_i = \frac{K_i - MK_i}{\sqrt{DK_i}}$.

4. Тепер рахуємо статистику $X^2 = \sum_{i=1}^k \bar{K}_i^2$.

Нульова гіпотеза приймається якщо $X^2 < \chi_\alpha^2(k-1)$, де $\chi_\alpha^2(k) - \alpha \cdot 100\%$ процентна точка χ^2 -розподілу з k степенями свободи.

Перевірка симетрій розподілу ранговими критеріями

Маємо ряд незалежних спостережень x_1, x_2, \dots, x_n над випадковою величиною ξ з функцією розподілу $F(x)$. Перевіримо симетричність розподілу відносно точки x_0 .

Гіпотеза: для дискретної випадкової величини $H_0 : F(x_0 + x) = 1 - F(x_0 - x + 0), \forall x$, або ж для неперервної випадкової величини $H_0 : p(x_0 + x) = p(x_0 - x), \forall x$.

Перевірка проводиться з деяким рівнем значимості α ($0 < \alpha < 1$).

Побудуємо послідовність z_1, z_2, \dots, z_n , де $z_i = |x_i - x_0|, i = \overline{1, n}$, а далі сформуємо варіаційний ряд $z_{(1)}, z_{(2)}, \dots, z_{(n)}$.

Абсолютним рангом виміру x_i називається порядковий номер значення $x_i = |x_i - x_0|$ у варіаційному ряді $z_{(1)}, z_{(2)}, \dots, z_{(n)}$, позначатимемо його як $R_{i,n}^+$ ($i = \overline{1, n}$).

Розіб'ємо вибірку x_1, x_2, \dots, x_n на дві вибірки, в першій всі виміри більше x_0 , в іншій решта. Позначення для індексів з першої вибірки $I^+ = \{i : x_i > x_0, i = \overline{1, n}\}$. Тепер можна порівняти дві наші вибірки на однорідність.

Аналог критерію нормальних міток

Статистика критерію має вигляд $C^+ = \sum_{i \in I^+} M^+(R_{i,n}^+, n)$, при справедливості H_0 :

$$MC^+ = \frac{n}{\sqrt{2\pi}}, \quad DC = \frac{1}{4} \sum_{i=1}^n (M^+(i, n))^2.$$

Аналог критерію Ван дер Вардена

Статистика критерію має вигляд $V^+ = \sum_{i \in I^+} \Phi^{-1} \left(\frac{1}{2} + \frac{R_{i,n}^+}{2(n+1)} \right)$, при справедливості H_0 :

$$MV^+ = \frac{1}{2} \sum_{i=1}^n \Phi^{-1} \left(\frac{1}{2} + \frac{i}{2(n+1)} \right), \quad DV^+ = \frac{1}{4} \sum_{i=1}^n \left(\Phi^{-1} \left(\frac{1}{2} + \frac{i}{2(n+1)} \right) \right)^2$$

Аналог критерію Вілкоксона

Статистика критерію має вигляд $S^+ = \sum_{i \in I^+} R_{i,n}^+$, при справедливості H_0 :

$$MS^+ = \frac{(n+1)n}{4}, \quad DS^+ = \frac{n(n+1)(2n+1)}{24}.$$

Нехай U^+ – одна з вищенаведених статистик. Стандартизуємо U^+ : $\bar{U}^+ = \frac{U^+ - NU^+}{\sqrt{DU^+}}$. Область прийняття гіпотези H_0 : $|\bar{U}^+| < u_{1-\alpha/2}$, де u_α – квантиль рівня α нормального розподілу з параметрами 0 та 1.

Визначення рангів у випадку наявності нерозрізних значень

Нехай $\nu_1, \nu_2, \dots, \nu_n$ – об’єднана вибірка, побудована на основі спостережень над змінними, що досліджуються. Існує два варіанти однакових спостережень:

1. спостереження стосуються однієї змінних, тоді використовується метод випадкового рангу: ранг однакових елементів є довільним числом, яке припало на цю множину значень.
2. спостереження стосуються різних змінних, тоді використовують або вже відомий нам метод випадкового рангу, або метод середньої мітки: всім рівним спостереженням присвоюють середнє значення мітки підраховане за множиною рангів, яка відповідає цій групі нерозрізними вимірів.

Корекція алгоритмів рангових критеріїв:

$DC = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^g \tau_i \bar{M}_i^2$ для критерію нормальних міток Фішера, $DV = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^g \tau_i (\bar{\Phi}_i^{-1})^2$ для критерію Ван дер Вардена, де g – кількість груп нерозрізними спостережень, τ_i – кількість значень у i -ій групі.

Зауваження! Метод середньої мітки потужніший методу випадкового рангу.

Видалення аномальних спостережень

Викиди – аномальні спостереження. До викидів будемо відносити ті виміри, значення яких не узгоджується з розподілом більшості аномальних спостережень.

Обробка скалярних вимірів. Існують наступні методи:

- критерій Граббса;
- критерій Томпсона;
- критерій Тітьєна-Мура;
- пробіт графік;
- ймовірнісний графік;
- зображення “стебло-листок”;
- зображення “скринька з вусами”

Критерії розглядаються для нормальних вибірок. Нехай маємо вибірку x_1, x_2, \dots, x_n . Гіпотеза H_0 : найбільш підозрілий на аномальність вимір не є викидом із рівнем значущості α ($0 < \alpha < 1$).

Критерій Граббса

Нагадаємо, що $\bar{x}(n) = \frac{1}{n} \sum_{i=1}^n x_i$, $s(n) = \sqrt{\frac{1}{n} (\sum_{i=1}^n x_i^2 - n\bar{x}^2(n))}$. Побудуємо послідовність z_1, z_2, \dots, z_n , де $z_i = |x_i - \bar{x}(n)|$, $i = \overline{1, n}$, та її варіаційний ряд $z_{(1)}, z_{(2)}, \dots, z_{(n)}$. Введемо допоміжне позначення $z_{(j)} = |x_{i(j)} - \bar{x}(n)|$, $j = \overline{1, n}$, тоді підозрливим на аномальність є елемент вибірки що відповідає останньому елементу $z_{(n)}$ варіаційного ряду, тобто $x_{i(n)}$. Розглянемо статистику

$$T(n) = \frac{x_{i(n)} - \bar{x}(n)}{s(n)}.$$

Областю прийняття гіпотези H_0 буде $|T(n)| < T_{\alpha/2}(n)$, де $T_{\alpha/2}(n) - 100\frac{\alpha}{2}\%$ точка розподілу статистики $\frac{x_{i(n)} - \bar{x}(n)}{s(n)}$. Якщо спостереження аномальне, то його викидають і так продовжується допоки останній елемент нової вибірки перестає бути аномальним.

Критерій Томпсона

Модифікація критерію Граббса із статистикою

$$t(n) = \frac{\sqrt{n-2} \cdot T(n)}{\sqrt{n-1-T^2(n)}}.$$

Область прийняття гіпотези $H_0: |t(n)| < t_{\alpha/2}(n-2)$, де $t_{\alpha/2}(n-2) - 100\frac{\alpha}{2}\%$ точка розподілу Стюдента з $(n-2)$ степенями свободи.

Критерій Тітьєна-Мура

Дозволяє перевіряти кілька спостережень на аномальність одразу. На базі варіаційного ряду з критерію Граббса обчислюємо статистику:

$$E(n, k) = \frac{\sum_{i=1}^{n-k} (z_{(i)} - \bar{z}(n-k))^2}{\sum_{i=1}^n (z_{(i)} - \bar{z}(n))^2},$$

де на аномальність перевіряються останні k членів варіаційного ряду, а $\bar{z}(m) = \frac{1}{m} \sum_{i=1}^m z_{(i)}$. Область прийняття гіпотези $H_0: E(n, k) \geq E_{1-\alpha}(n, k)$, де $E_{1-\alpha}$ – квантиль рівня α розподілу статистики $E(n, k)$.

Випадок векторних значень

Нехай маємо векторні величини x_1, x_2, \dots, x_n , $x_i \in \mathbb{R}^q$, $i = \overline{1, n}$. Гіпотеза H_0 : найбільш підозрілий на аномальність вектор не є викидом із рівнем значимості α ($0 < \alpha < 1$).

Критерій на базі F -статистики

Введемо величини $\bar{x}_i = \frac{1}{n-1} \sum_{j \neq i} x_j$, $\widehat{\sum}_i = \frac{1}{n-2} \sum_{j \neq i} (x_j - \bar{x}_i)(x_j - \bar{x}_i)^T$, $i = \overline{1, n}$. Обчислимо вибіркові відстані Махаланобіса:

$$D_i^2 = (x_i - \bar{x}_i)^T \cdot \widehat{\sum}_i^{-1} (x_i - \bar{x}_i), \quad i = \overline{1, n}.$$

Визначимо наступну статистику:

$$F_i = \frac{(n-1)(n-1-q)}{n(n-2)q} D_i^2.$$

Знаходимо індекс i_0 найбільш підозрілого вектора як $i_0 = \arg \max_i F_i$. Гіпотеза H_0 приймається як $F_{i_0} < F_{\alpha}(q, n-1-q)$, де $F_{\alpha}(q, n-1-q) - 100\alpha\%$ точка F -розподілу з q та $n-1-q$ степенями свободи.

Розвідувальний аналіз

Один з елементів попередньої обробки даних. Для аналізу над однією змінною можна використовувати:

- пробіт-графік (англ. probit plot);
- ймовірнісний графік (probability plot);
- висячі гістобари (hanging histobars);
- підвішену коренеграму (suspended rootgram);
- зображення “скринька з вусами” (boxplot).

Сім’ї розподілів типу зсув-масштабу

Перші два графічних представлення даних можуть використовуватися для різних функцій розподілів, що належать до сімей типу зсув-масштабу. Сім’я розподілів \mathcal{F} називається сім’єю розподілів типу зсув-масштабу, якщо існує така базова функція розподілу $F_0(\cdot) \in \mathcal{F}$, що для довільної функції розподілу $F(\cdot)$ з цієї сім’ї існують дійсні a та $b > 0$ такі, що $F(x) = F_0\left(\frac{x-a}{b}\right)$.

Прикладами сімей типу зсув-масштабу можуть слугувати:

1. Сім’я нормальних розподілів. Справді, якщо $\xi \sim N(m, \sigma^2)$, то $F_\xi(x) = \Phi\left(\frac{x-m}{\sigma}\right)$, де $\Phi(\cdot)$ – функція розподілу випадкової величини з нормальним розподілом з параметрами 0 та 1.
2. Сім’я показникових (експоненціальних) розподілів. справді, якщо $\xi \sim \text{Exp}(\lambda)$, то $F_\xi(x) = \Phi(\lambda x)$, де $\Phi(\cdot)$ – функція розподілу випадкової величини з показниковим розподілом з параметром 1.

Пробіт графік

Нехай маємо вибірку об’єму n спостережень x_1, x_2, \dots, x_n над скалярною змінною ξ з функцією розподілу $F_\xi(x)$. За вибіркою обчислюємо емпіричну функцію розподілу $\hat{F}_\xi(x)$. Пробіт-графік деякої сім’ї розподілів \mathcal{F} типу зсув-масштабу з базовою функцією F_0 задається таким чином: $y = F_0^{-1}(\hat{F}_\xi(x))$.

Відзначимо, що якщо $F_\xi \in \mathcal{F}$, то пробіт-графік буде прямою (а точніше буде близьким до неї адже $F_\xi \approx \hat{F}_\xi$, але взагалі кажучи не факт що $F_\xi = \hat{F}_\xi$).

Також пробіт-графік дозволяє виявляти викиди: якщо більшість точок узгоджується з певним законом розподілу (лежать на прямій), а якась одна ні (лежить осторонь), то цілком можливо що вона є аномальним спостереженням.

Ймовірнісний графік

Нехай $\hat{F}_\xi(x)$ – емпірична функція розподілу, обчислена за спостереженнями x_1, x_2, \dots, x_n над скалярною випадковою величиною ξ . Ймовірнісний графік це графік функції $y = \hat{F}_\xi(x)$ побудований на ймовірнісному папері деякої сім’ї розподілів типу зсув-масштабу з базовою функцією $F_0(\cdot)$. Він відрізняється від звичайного паперу тим, що смуга $0 \leq y \leq 1$ трансформується за правилом $(x, y) \mapsto (x, F_0^{-1}(y))$.

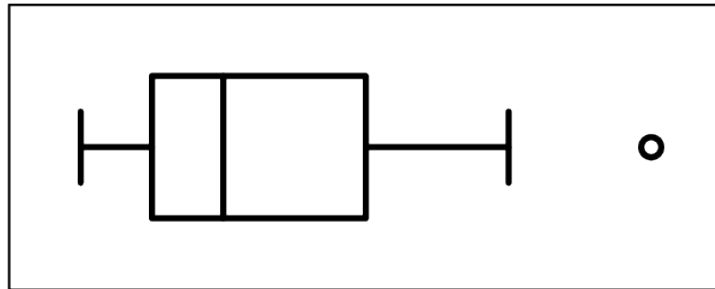
Висячі гістобари

За вибіркою обчислюємо \hat{m} , $\hat{\sigma}^2$, будемо графік щільності $N(\hat{m}, \hat{\sigma}^2)$, і у центрі кожного інтервалу групування даних підвішуємо до точки на графіку щільності гістобару (стовпчик), довжина якого пропорційна відносній частоті потрапляння виміру у цей інтервал. Гіпотезу про нормальність розподілу приймають якщо гістобара не сильно відхиляється від осі абсцис.

Підвішена коренеграма

Являє собою послідовність прямокутників, побудованих у центрах інтервалів групування даних вибірки x_1, x_2, \dots, x_n , довжини який пропорційні різниці $\sqrt{\nu} - \sqrt{\nu_0}$, де ν і ν_0 – емпірична і теоретична для нормального розподілу відповідно частоти потрапляння у інтервал групування.

Зображення “скринька з вусами”



- Середня лінія – медіана.
- Ліва границя – нижній кuartиль.
- Права границя – верхній кuartиль.
- Лівий вус – найменше значення у вибірці.
- Правий вус – найбільше значення у вибірці.
- Викиди позначаються як окремі точки поза вусами.