Попередня обробка даних

Попередня обробка даних займається знаходженням по спостереженням змінних їх первинних характеристик та допоміжної інформації, які стануть у нагоді у подальших розділах обробки та аналізу інформації (даних).

До попередньої обробки даних входять:

- *підрахування базових характеристик розподілу змінної*, яка спостерігається (а саме: початкових та центральних моментів, квантилів та відсоткових точок, характеристик центру значень змінної, характеристик розсіювання значень змінної, асиметрії, ексцесу, емпіричної функції розподілу, емпіричної функції щільності і т.п.),
- *виявлення та вилучення аномальних спостережень*,
- *перевірка основних гіпотез* (а саме: стохастичності вибірки, симетрії розподілу, згоди з заданим законом розподілу, однорідності вибірки, і т.п.),
- *розвідувальний аналіз* (проводить візуальний попередній експрес-аналіз інформації).

Базові характеристики

До них відносять: початкові та центральні моменти, квантилі та відсоткові точки, характеристики центру значень змінної, характеристики розсіювання значень змінної, асиметрію, ексцес, емпіричну функцію розподілу, емпіричну функцію щільності та інші характеристики, які будуть служити хорошим підґрунтям для подальшого аналізу даних.

Згадаємо деякі поняття.

Означення. ***Варіаційним рядом*** вибірки

$$x_1, x_2, \ldots, x_n$$

називається така послідовність

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)},$$

яка утворена з цієї вибірки після розташування її значень у порядку не спадання, тобто справедливо

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}.$$

Означення. Члени варіаційного ряду

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

називаються ***порядковими статистиками***.

Означення. i – ий член варіаційного ряду $x_{(i)}$ називається i – ***ою порядковою статистикою.***

Квантилі та відсоткові точки розподілу

Ці поняттями будуть широко використовуватися у подальшому при розв'язанні задач перевірки гіпотез, побудові довірчих інтервалів та областей і т.п. Визначати їх будемо окремо для неперервних та дискретних розподілів. Спочатку дамо означення для теоретичних значень квантилів.

Означення. ***[Теоретичним] квантилем рівня*** q ***розподілу*** (q-квантилем розподілу) ***неперервної випадкової величини*** ξ називається таке дійсне число u_q, яке визначається з рівняння

$$P\{\xi < u_q\} = q, \quad 0 < q < 1.$$

Означення. ***[Теоретичним] квантилем рівня*** q ***розподілу*** (q-квантилем розподілу) ***неперервної випадкової величини*** ξ називається таке дійсне число u_q, яке визначається з рівняння

$$P\{\xi < u_q\} = q, \quad 0 < q < 1.$$

Означення. ***[Теоретичним] квантилем рівня*** q ***розподілу*** (q-квантилем розподілу) ***дискретної випадкової величини*** ξ з варіаційним рядом своїх значень $\{y_{(i)}\}$ називається довільне значення u_q з інтервалу $\left(y_{(i(q))}, y_{(i(q)+1)}\right]$, для границь якого справедливо

$$P\{\xi < y_{(i(q))}\} < q,$$

$$P\{\xi < y_{(i(q)+1)}\} \ge q, \quad (0 < q < 1).$$

Означення. ***Емпіричним (вибірковим) квантилем рівня*** q ***розподілу*** випадкової величини ξ називається квантиль рівня q відповідного емпіричного (вибіркового) розподілу.

Приклади квантилів.

1.***Медіана*** – це квантиль рівня 0,5, тобто $u_{0,5}$.

2.***Нижній та верхній квартилі*** визначаються як $u_{0,25}$ та $u_{0,75}$ відповідно.

3.***Децилі*** – це квантилі $\left\{u_{\frac{i}{10}}\right\}_{i=1}^{9}$.

4.***Центилі*** задаються наступним чином $\left\{u_{\frac{i}{100}}\right\}_{i=1}^{99}$.

5.***Інтерквантильна широта рівня*** q $\left(0<q<\frac{1}{2}\right)$ – це величина, яка обчислюється по формулі

$$\left(u_{1-q}-u_q\right).$$

6.***Інтерквартильна широта*** – це інтерквантильна широта рівня 0.25, а саме

$$\left(u_{0,75}-u_{0,25}\right).$$

7.***Ймовірнісне відхилення*** d_ξ визначається як половина інтерквартильної широти, тобто $d_\xi=\frac{1}{2}\left(u_{0,75}-u_{0,25}\right)$.

8.***Інтерсектильна широта*** – це інтерквантильна широта рівня $\frac{1}{6}$, тобто

$$\left(u_{\frac{5}{6}}-u_{\frac{1}{6}}\right).$$

9.***Інтердецильна широта*** – це інтерквантильна широта рівня 0.1, а саме $\left(u_{0,9}-u_{0,1}\right)$.

Означення. ***[Теоретичною]*** *Q-відсотковою точкою розподілу* ***неперервної випадкової величини*** ξ називається таке дійсне число v_Q, яке є розв'язком рівняння

$$P\{\xi \geq v_Q\} = \frac{Q}{100}, \quad 0 < Q < 100.$$

Означення. ***[Теоретичною] Q-відсотковою точкою розподілу дискретної випадкової величини*** ξ з варіаційним рядом своїх значень $\{y_{(i)}\}$ називається довільне значення v_Q з інтервалу $\left(y_{(i(Q))}, y_{(i(Q)+1)}\right]$, для границь якого справедливо

$$P\{\xi \geq y_{(i(Q))}\} > \frac{Q}{100},$$

$$P\{\xi \geq y_{(i(Q)+1)}\} \leq \frac{Q}{100}, \quad 0 < Q < 100.$$

Означення. ***Емпіричною (вибірковою)*** *Q-відсотковою точкою* ***розподілу*** випадкової величини ξ називається Q-відсоткова точка відповідного емпіричного (вибіркового) розподілу.

Ці два поняття взаємно доповнюють одне одного. У неперервному випадку для певного розподілу взаємозв'язок між ними прозорий і має наступний вигляд:

$$u_q = v_{(1-q)100}, \quad v_Q = u_{1-\frac{Q}{100}}.$$

Для широко вживаних розподілів складені відповідні таблиці, з яких легко визначити потрібні квантилі та відсоткові точки.

Самостійна робота №2. З навчального посібника «Слабоспицький О.С. Аналіз даних. Попередня обробка, 2001» необхідно пропрацювати матеріал наведений у **розділах 2.1-2.5**:

Характеристики положення центра значень змінної.

Характеристики розсіювання значень змінної.

Аналіз скошеності та гостроверхості розподілу.

Характеристики випадкових векторів.

Виявлення та вилучення аномальних спостережень

Причини появи у вибірках аномальних спостережень (викидів): збій у роботі обладнання, суттєві похибки вимірювальних приладів, порушення умов проведення експерименту, стихійні лиха, форс-мажорні обставини, інші непередбачувані причини і т.п.

Означення. ***Аномальними спостереженнями (викидами)*** у вибірці називаються ті виміри з неї, значення яких не узгоджуються з розподілом більшості отриманих спостережень.

Виявлені у вибірці аномальні спостереження, при наявності можливості *корегують*, інакше, як правило, їх *видаляють* з вибірки, *але обережно*. Бо головне при цьому «не вихлюпнути разом з водою і дитя», бо саме екстремальні виміри, які у більшості випадків і підозрюють на аномальність, дуже часто є найбільш інформативними спостереженнями. Саме спостереження отримані *під час функціонування досліджуємого об'єкту в екстремальних режимах*, як правило, і несуть найбільше інформації про цей об'єкт.

Приклад 2. Випробування літальних апаратів. А саме …

Так як найбільш розробленою виявилася теорія *для нормальних вибірок*, то саме її і будемо розглядати у подальшому.

I. ***Випадок обробки скалярних спостережень.*** Нехай маємо

$$\xi: x_1, x_2, \ldots, x_n.$$

Потрібно виявити та вилучити викиди з цієї вибірки. Для цього було запропоновано ряд методів, а саме:

1) критерій Граббса,
2) критерій Томпсона,
3) критерій Тітьєна-Мура,
4) засоби розвідувального аналізу:
   - пробіт-графік,
   - ймовірнісний графік,
   - зображення "скринька з вусами",
   - зображення "стебло-листок".

Зупинимося на кожному з них більш детальніше.

Зауваження. У подальшому при перевірці деякої гіпотези H_0, якщо не буде згадуватися протилежна гіпотеза H_1, то по замовчуванню будемо вважати, що вона є простим запереченням H_0.

1) Критерій Граббса. Тут задачу розв'язують шляхом перевірки:
H_0 : *найбільш підозрілий на аномальність вимір не є викидом*, $\alpha > 0$.

1. Спочатку за вибіркою підрахуємо такі статистики:

$$\overline{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s(n) = \sqrt{\frac{1}{n}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2(n)\right]}.$$

2. Далі будуємо послідовність:

$$z_1, z_2, \quad \ldots, \quad z_n,$$

де $z_i = \left|x_i - \overline{x}(n)\right|$, $i = \overline{1,n}$.

3. А потім відповідний варіаційний ряд:

$$z_{(1)}, z_{(2)}, \quad \ldots, \quad z_{(n)},$$

де $z_{(j)} = \left|x_{i(j)} - \overline{x}(n)\right|$, $j = \overline{1,n}$.

Підозрюємо на аномальність спостереження, яке відповідає останньому члену варіаційного ряду $z_{(n)}$, а саме $x_{i(n)}$. Перевіримо його на аномальність.

4. Для цього обчислимо таку статистику:

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

5. Тоді логічно за *область прийняття гіпотези* H_0, що $x_{i(n)}$ не є викидом, взяти область, яка не включає у себе екстремальних значень останньої статистики, а саме:

$$\left|T(n)\right| < T_{\frac{\alpha}{2}}(n),$$

де $T_{\frac{\alpha}{2}}(n)$ - $100\dfrac{\alpha}{2}$ відсоткова точка розподілу статистики $\dfrac{x_{i(n)} - \overline{x}(n)}{s(n)}$.

Якщо $x_{i(n)}$ є аномальним, то його видаляють з вибірки і всю процедуру повторюють починаючи з пункту 1, але вже з отриманою

скороченою вибіркою. Все це повторюється до тих пір, доки на деякому кроці найбільш підозрілий на аномальність вимір виявиться не викидом. Після цього роботу завершують.

Зауваження 1. Відповідна таблиця цих відсоткових точок отримана Граббсом і наведена у довіднику:
Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.

Зауваження 2. Критерій Граббса, в основному, використовується для невеликих значень n.

2) Критерій Томпсона. Він є модифікацією критерію Грабса. Знову задача розв'язується шляхом перевірки гіпотези:
H_0 : *найбільш підозрілий на аномальність вимір не є викидом*, $\alpha > 0$.

1. ……
2. ……
3. ……
4. Для цього обчислимо таку статистику:

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

5. А далі будується статистика:

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

розподіл якої можна наблизити t – розподілом Стьюдента з $(n-2)$ степенями свободи при достатньо великих n.

6. Аналогічно як у попередньому критерії область прийняття нашої гіпотези набуває вигляду:

$$|t(n)| < t_{\frac{\alpha}{2}}(n-2),$$

де $t_{\frac{\alpha}{2}}(n-2)$ – $100\dfrac{\alpha}{2}$ відсоткова точка t-розподілу Стьюдента з $(n-2)$ степенями свободи, використання якої є більш зручнішим.

А далі, якщо вимір $x_{i(n)}$ виявився аномальним, то його вилучають з вибірки і всю процедуру повторюють з пункту 1, але вже зі скороченою вибіркою, до тих пір поки найбільш підозрілий на аномальність вимір виявиться не викидом.

Самостійна робота №3. З навчального посібника «Слабоспицький О.С. Аналіз даних. Попередня обробка, 2001» необхідно пропрацювати матеріал наведений у:
Додаток 1. Нормальний закон та пов'язані з ним розподіли.

3) Критерій Тітьєна-Мура. Є можливість перевірки на аномальність одразу $k\,(k \geq 1)$ найбільш підозрілих спостережень. Розв'язання такої задачі зводиться до перевірки гіпотези:

H_0 : *не усі k вимірів, найбільш підозрілі на аномальність, є викидами*, $\alpha > 0$.

---

1. Знову підраховуємо $\overline{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2. Далі будуємо послідовність:

$$z_1, z_2, \quad \ldots, \quad z_n,$$

де $z_i = \left| x_i - \overline{x}(n) \right|,\ i = \overline{1,n}$.

3. А потім відповідний варіаційний ряд:

$$z_{(1)}, z_{(2)}, \quad \ldots, \quad z_{(n-k)}, z_{(n-k+1)}, \quad \ldots, z_{(n)},$$

де $z_{(j)} = \left| x_{i(j)} - \overline{x}(n) \right|,\ j = \overline{1,n}$.

Найбільш підозрілими на аномальність будемо вважати k вимірів $x_{i(j)}$, які відповідають k останнім членам варіаційного ряду $z_{(1)}, z_{(2)}, \quad \ldots, \quad z_{(n)}$, а саме виміри: $x_{i(n-k+1)}, x_{i(n-k+2)}, \ldots, x_{i(n)}$.

4. Тепер обчислимо статистику:

$$E(n,k)=\frac{\sum_{i=1}^{n-k}\left(z_{(i)}-\overline{z}(n-k)\right)^2}{\sum_{i=1}^{n}\left(z_{(i)}-\overline{z}(n)\right)^2},\quad \text{де } \overline{z}(m)=\frac{1}{m}\sum_{i=1}^{m}z_{(i)}.$$

5. Логічно в якості *області прийняття нашої гіпотези* H_0, що спостереження $x_{i(n-k+1)}, x_{i(n-k+2)}, \ldots, x_{i(n)}$ не є викидами, взяти область, яка не включає у себе область малих значень статистики $E(n,k)$:

$$E(n,k) > E_\alpha(n,k),$$

де $E_\alpha(n,k)$ - квантиль рівня α розподілу статистики $E(n,k)$.

Якщо спостереження $x_{i(n-k+1)}, x_{i(n-k+2)}, \ldots, x_{i(n)}$ виявилися аномальними, то їх видаляють з вибірки і всю процедуру повторюють, але вже зі скороченою вибіркою, інакше алгоритм завершує свою роботу.

Недоліки критерію Тітьєна-Мура.

- не формалізована процедура вибору значення $k\,(k \ge 1)$ (можна запропонувати метод ділення навпіл: $k := k_0$; а потім $k := [k/2]$ тобто в результаті k буде приймати таку послідовність значень $k_0, [k_0/2], [k_0/2^2], \ldots, 3, 2, 1$ (тут використовується критерій Граббса або Томпсона), де k_0 - стартове значення),
- критерій Тітьєна-Мура сильно залежить від нормальності.

4) Засоби розвідувального аналізу видалення аномальних скалярних спостережень розглядаються у відповідному розділі, який присвячений розвідувальному аналізу.

II. *Випадок обробки векторних спостережень.*