УДК 519.2
ББК 22.172я73
С 48

О.С.Слабоспицький

С 48 Аналіз даних. Попередня обробка: Навчальний посібник. – К.: Видавничо-поліграфічний центр "Київський університет", 2001. – 52 с.

Рецензенти: І.В.Бейко, д-р техн. наук, проф.;
В.А.Заславський, канд. фіз.-мат. наук, доц.

Посібник присвячено методам обробки та аналізу інформації. Основна увага приділяється задачам, які виникають на етапі її попередньої обробки. Описуються можливості використання розвідувального аналізу для цих цілей. Робота буде корисна студентам при вивченні курсу аналізу даних.

Рекомендовано до друку
Вченою Радою
факультету кібернетики
25 червня 2001 року

Вступ

Потоки інформації, які отримуються нами повсякденно нічого не варті, якщо ми не в стані провести їх ефективну обробку та зробити правильні висновки з її результатів. Саме тут нам у нагоді можуть стати математичні методи аналізу даних, які й дозволяють розв'язувати вищезгадані задачі. Широкому застосуванню цього арсеналу засобів також сприяла розробка відповідних програмних продуктів для персональних комп'ютерів, які набагато прискорили впровадження аналізу даних при проведенні досліджень у науці, економіці, бізнесі, медицині, біології і т.п. Вони показали свою ефективність при проведенні не тільки науково-дослідних робіт, але й прикладних соціологічних, маркетингових та різного роду аналітичних досліджень.

Можна виділити наступні етапи аналізу даних: отримання вхідної інформації, безпосередньо сама обробка її, аналіз та інтерпретація результатів обробки даних. Отримати результати обробки інформації це ще не все. Головне зробити правильні висновки з них. І саме це є принциповим, визначальним моментом.

Потрібно також підкреслити, що аналіз даних охоплює обробку, як кількісних так і якісних даних. Причому, можуть використовуватися не обов'язково ймовірнісні моделі при описі об'єктів, явищ, процесів, які досліджуються. Це суттєво розширює як коло можливостей самого аналізу даних так і області його застосування.

До розділів аналізу даних, які знайшли свос широке застосування, можна віднести наступні: попередня обробка даних, кореляційний аналіз, дисперсійний аналіз, регресійний аналіз, коваріаційний аналіз, дискримінантний аналіз, кластерний аналіз, аналіз часових рядів. Кожна з цих частин курсу відповідальна за розв'язок певного кола задач по обробці даних, які досліджуються, і має в своєму розпорядженні солідний арсенал математичних методів.

Робота, яка пропонується, присвячена висвітленню можливостей, які нам надає попередня обробка даних. Приділяється також увага популярному зараз розвідувальному аналізу даних.

Автор щиро вдячний студентам факультету кібернетики Київського національного університету імені Шевченка, які всіляко сприяли покращанню даного посібника.

Всі зауваження та побажання по даній роботі будуть з вдячністю сприйняті автором і можуть бути надіслані електронною поштою (e-mail: *sl@planetmail.com*).

1. Опис та підготовка вхідної інформації

Перш ніж проводити будь-яку обробку даних потрібно з'ясувати спочатку з яким типом даних маємо справу і дати їх формальний опис. У разі потреби, перейти до іншого представлення отриманої інформації. Далі буде дана характеристика можливих типів вхідної інформації та проведено знайомство з процедурою групування даних.

1.1. Класифікація змінних

Почнемо з того, що з'ясуємо, якого роду інформація може поступати на вхід для обробки, аналізу та інтерпретації. Перш за все потрібно сказати, що значення змінних, які спостерігаються, можуть бути як *кількісні* так і *якісні*.

З кількісними змінними наше знайомство відбулося настільки давно, що питань по цьому класу не повинно виникати. А от із якісними змінними познайомимося поближче.

Якісні змінні поділяють на *ординальні* та *номінальні*. Ординальні змінні ще називають *порядковими*, а номінальні – *класифікаційними*. Що спільного і в чому відмінність цих змінних. Як ординальні так і номінальні змінні приймають свої значення з деякої множини, елементи якої називають *градаціями*. Відмінність полягає в тому, що градації, які приймає як свої значення ординальна змінна, природно впорядковані по степені прояву властивості, яку представляє ця ординальна змінна. А градації, які приймає як свої значення номінальна змінна, такого порядку не мають.

Прикладом номінальної змінної може служити змінна, яка приймає як свої значення назви предметів, що вивчаються студентами: математичний аналіз, алгебра, програмування і т.д. Навчаючись на факультеті студенти можуть отримати диплом бакалавра, спеціаліста або магістра. Змінна, яка приймає ці значення, може служити прикладом ординальної змінної, бо ці градації природно впорядковані по рівню отриманої освіти.

Крім цього серед якісних змінних виділяють *категорізовані* та *некатегорізовані*. До категорізованих змінних відносять змінні, для яких повністю визначена множина градацій та правило віднесення

значення змінної, яке спостерігається, до певної градації. У протилежному випадку змінну називають некатегорізованою.

Остання класифікація змінних зовсім прозора. Це поділ змінних на *дискретні* та *неперервні*.

1.2. Групування даних

Нехай у нашому розпорядженні для скалярної змінної ξ, яка досліджується, мається вибірка об'єму n: $x_1, x_2, \ldots, x_n$.

Якщо вибірка невелика, то можна безпосередньо здійснити її обробку. 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. Для цього у ряді випадків використовують *групування 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 можна здійснити наступним чином. Визначивши екстремальні значення вибірки $x_{\min} = \min_i x_i$ та $x_{\max} = \max_i x_i$, інтервал $[x_{\min}, x_{\max}]$ розбивають на s однакових підінтервалів $[a_i, b_i)$, $i = \overline{1,s}$. Праву границю кожного (крім останнього) з підінтервалів, для однозначності, будемо включати у наступний підінтервал. Значення s, як правило, вибирають не меншим 5, але не більшим 30, залежно від значення об'єму вибірки n. Для наближеного підрахування значення s, яке не повинно виходити за вищевказані межі, можна використовувати наступні формули $1+\lfloor \log_2 n \rfloor$ або $\lfloor 10\lg(n) \rfloor$. Після цього для кожного з підінтервалів $[a_i, b_i)$ $\left(i = \overline{1,s}\right)$ визначають значення його центральної точки $x_i^* = \frac{a_i + b_i}{2}$ та кількість вимірів v_i з нашої вибірки, які потрапили в цей підінтер-

вал. Таким чином, здійснено перехід від вибірки $x_1, x_2, \ldots, x_n$ до набору значень $\{x_i^*, \nu_i\}_{i=1}^s$ ($\sum_{i=1}^s \nu_i = n$), з яким у подальшому і оперують.

Зазначимо, що при проведенні групування даних зовсім не обов'язково брати підінтервали однакової довжини. Відмова від цього обмеження може дати більш кращі результати обробки інформації.

Для прикладу продемонструємо як відбувається побудова емпіричних функцій розподілу та щільності на базі згрупованих даних.

Надалі будемо використовувати наступні позначення для характеристик випадкової величини ξ: $F_\xi(x) = P\{\xi < x\}$ – *функція розподілу*, $p_\xi(x)$ – *функція щільності*, $\{y_i, p_i\}_{i=1}^m$ – *полігон ймовірностей*, у випадку коли ξ є дискретною випадковою величиною, яка набуває значення y_i з ймовірностями p_i, $i = \overline{1,m}$.

Подивимося, який вигляд будуть мати відповідні оцінки цих характеристик, що побудовані по згрупованим даним $\{x_i^*, \nu_i\}_{i=1}^s$. *Емпірична (вибіркова) функція розподілу* $\hat{F}_\xi(x)$ буде визначатися наступним чином:

$$\hat{F}_\xi(x) = \frac{1}{n} \sum_{i:\, b_i \le x} \nu_i \ .$$

У свою чергу *емпірична (вибіркова) функція щільності* $p_\xi(x)$ підраховується за формулою:

$$\hat{p}_\xi(x) = \frac{\nu_{i(x)}}{n(b_{i(x)} - a_{i(x)})},$$

де $i(x)$ – номер підінтервалу якому належить x.

1.3. Моделювання змінних

Під час роботи з даними виникає потреба в генерації спостережень над випадковими величинами з заданими функціями розподілу. Отримати такі вибірки можна шляхом моделювання потрібних значень. Один із підходів при розв'язанні цієї задачі полягає у пред-

ставленні величини ξ, яка моделюється, у вигляді деякої функції $g(\xi_1, \xi_2, \ldots, \xi_q)$ від найпростіших випадкових величин $\xi_1, \xi_2, \ldots, \xi_q$. Як правило, у цій ролі виступають незалежні $\xi_1, \xi_2, \ldots, \xi_q$, рівномірно розподілені на відрізку $[0,1)$. Тому, задача зводиться до необхідності вміти розв'язувати наступні дві проблеми:

- моделювання незалежних $\xi_1, \xi_2, \ldots, \xi_q$, рівномірно розподілених на відрізку $[0,1)$,
- знаходження потрібної функції $g(\cdot, \cdot, \ldots, \cdot)$ у представленні величини ξ, яку потрібно моделювати.

Нижче кожну з них розглянемо окремо.

Перша проблема розв'язується шляхом використання *датчика (генератора) випадкових чисел* – спеціального пристрою, який після запиту на виході дозволяє отримати реалізацію випадкової величини з заданим законом розподілу. Повторні звернення до цього генератора дозволяють отримати наступні незалежні спостереження над цією випадковою величиною. Найбільший інтерес для нас представляє датчик рівномірно розподіленої на відрізку $[0,1)$ випадкової величини. Познайомимося з генераторами випадкових чисел детальніше.

1.3.1. Класифікація датчиків випадкових чисел

Виділяють наступні *класи датчиків* (генераторів) *випадкових чисел*:

- табличні,
- фізичні,
- програмні.

Охарактеризуємо кожен із цих типів генераторів окремо.

Табличний датчик випадкових чисел являє собою таблицю заповнену реалізаціями випадкової величини з заданим законом розподілу. Представлені у таких таблицях вибірки, як правило, досить високої якості, але вони мають обмежений об'єм. Та й кількість таких вибірок невелика. Це суттєво стримує їх використання.

Фізичний датчик випадкових чисел конструюється на основі деякого електронного пристрою, на виході якого спостерігають потрібну реалізацію. Ці генератори надають можливість отримувати ви-

бірки довільного об'єму, що не дозволяли робити табличні датчики. Але вони мають інший недолік – кожна отримана вибірка є унікальною і повторити її практично неможливо. Цього недоліку позбавлений наступний клас генераторів.

Програмний датчик випадкових чисел будується на базі деякої програми, на виході якої формується потрібна реалізація. Ці програми, як правило, базуються на використанні певних рекурентних формул із деякою глибиною пам'яті. Задаючи у рекурентному співвідношенні однакові початкові значення (стартові числа), можна повторити конкретну вибірку довільну кількість раз. Але ці генератори, в свою чергу, мають свій недолік – вони періодичні. В принципі, числа, які отримуються на виході, потрібно називати "псевдовипадковими" числами, бо вони формуються згідно з детермінованого закону, а саме деякою рекурентною формулою.

1.3.2. Програмні датчики та їх властивості

В основі програмних генераторів, як правило, лежить використання рекурентних формул. Саме вибір їх конкретного вигляду і буде визначати властивості цих датчиків. Познайомимося з деякими лінійними, а потім нелінійними формулами, які знайшли застосування в генераторах рівномірно розподілених на відрізку $[0,1)$ випадкових величин.

Широкого розповсюдження при побудові цих датчиків набула *лінійна змішана формула*. Вона має наступний загальний вигляд:

$$\begin{cases} x_i = \dfrac{\widetilde{x}_i}{M}, \\ \widetilde{x}_i = \left(a_0 + \sum\limits_{j=1}^{l} a_j \widetilde{x}_{i-j} \right) \bmod M, \quad i = 1, 2, \ldots \end{cases},$$

де всі параметри алгоритму є цілими і задовольняють таким умовам: $l \geq 1$, $a_j \geq 0$ $(j = \overline{0,l})$, $M > 0$, а стартові числа лежать у межах $0 \leq \widetilde{x}_{i-j} \leq M-1$, $(j = \overline{1,l})$.

Так як по побудові послідовність $\{\widetilde{x}_i\}_{i>0}$ набуває значення з множини $\{0, 1, \ldots, M-1\}$, то послідовність $\{x_i\}_{i\geq 0}$ в свою чергу буде набувати значення з потрібного інтервалу $[0,1)$. Зауважимо, що значення $\{\widetilde{x}_i\}_{i\geq 0}$ можна використовувати при моделюванні рівномірного розподілу на множині значень $\{0, 1, \ldots, M-1\}$.

Аналіз лінійної змішаної формули почнемо з деяких її частинних випадків. Розглянемо випадок, коли $a_0 = 0$ та $l = 1$. Кажуть, що у цьому випадку використовується *мультиплікативний конгруентний метод*, а сам алгоритм набуває вигляд:

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

Для того, щоб скористатися цією процедурою достатньо знати одне стартове число $\widetilde{x}_0$ $(0 \leq \widetilde{x}_0 \leq M-1)$. Очевидно, що вибір $\widetilde{x}_0 = 0$ буде невдалим, бо тоді вся послідовність $\{\widetilde{x}_i\}_{i\geq 0}$ буде тотожна нулеві і така вибірка не представляє цінності.

З алгоритму випливає, що можливими значеннями у послідовності $\{\widetilde{x}_i\}_{i\geq 0}$ є значення з множини $\{0, 1, \ldots, M-1\}$, а сама вона буде періодичною. Причому максимальне значення періоду T_{max} не буде перевищувати M, а це в свою чергу примушує вибирати M якомога ближчим до максимального цілого, що допускається на конкретному комп'ютері, і яке будемо позначати у подальшому через $\max\text{int}$. Наприклад, як M можемо взяти найбільше просте число, яке менше $\max\text{int}$. Таким чином, виникає потреба у виборі параметрів алгоритму мультиплікативного конгруентного методу так, щоб максимізувати його період. А так як період довжини M повинен містити у собі значення рівне нулеві, то це приводить до подальшого виродження послідовності $\{\widetilde{x}_i\}_{i\geq 0}$ у нуль, що приводить до висновку, що мультиплікативний конгруентний метод не дозволяє досягти максимального теоретично можливого періоду рівного M.

Визначимо функцію $\lambda(M)$ наступним чином:

$$\lambda(M)=\begin{cases}1, & \text{якщо } M=2,\\ 2, & \text{якщо } M=4,\\ p^{q-1}(p-1), & \text{якщо } M=p^q \quad (q\ge 1, \text{просте } p>2),\\ НСК(\lambda(p_1^{q_1}),\lambda(p_2^{q_2}),\dots,\lambda(p_k^{q_k})), & \text{якщо } M=p_1^{q_1}p_2^{q_2}\cdots p_k^{q_k}\\ & (q_i\ge 1, \text{прості } p_i>2, i=\overline{1,k}),\end{cases}$$

де $НСК(n_1,n_2,\dots,n_k)$ – найменше спільне кратне для позитивних цілих чисел $n_1,n_2,\dots,n_k$.

Наступне відоме твердження дозволяє з'ясувати питання про максимально можливий період досліджуємого методу.

Теорема 1. Максимальний період послідовності $\{\widetilde{x}_i\}_{i\ge 0}$ мультиплікативного конгруентного методу $T_{\max}=\lambda(M)$. Для того, щоб він досягався достатньо виконання наступних умов:

1) $\widetilde{x}_0$ та M є взаємно прості числа,

2) $a_1^{\lambda(M)} \bmod M=1$, тобто a_1 є первісним елементом по модулю M.

Зауваження. Якщо покласти M рівним деякому простому числу, то $T_{\max}=M-1$. Тобто у цьому випадку період буде тільки на одиницю менший від M. Наприклад, можна взяти найбільше ціле, яке менше maxint. Тоді залежно від розрядності комп'ютера q можна скористатися наступними значеннями для M:

q	M
16	2^{16}-15
32	2^{32}-5
64	2^{64}-59

А використання у ролі M значення 2^q $(q\ge 3)$ дозволяє досягти лише періоду $M/4$.

Перейдемо тепер до розгляду випадку, коли $a_0>0$ та $p=1$. Метод, який йому відповідає, називається *змішаним конгруентним методом*, а безпосередньо алгоритм має вигляд:

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

Вибір параметрів цієї процедури з метою досягнення максимального періоду можна здійснити скориставшись нижченаведеним твердженням.

Теорема 2. Для того, щоб визначена згідно змішаного конгруентного методу послідовність $\{\widetilde{x}_i\}_{i\ge 0}$ досягала свого максимального періоду $T_{\max}=M$ необхідно і достатньо, щоб виконувалися наступні умови:

1) a_0 і M – взаємно прості,

2) $(a_1-1) \bmod p=0$ для кожного простого p, яке є дільником M,

3) $(a_1-1) \bmod 4=0$, якщо $M \bmod 4=0$.

Зауваження. Вибір параметрів алгоритму змішаного конгруентного методу, які забезпечують його максимальний період, не може бути гарантією високої якості побудованого датчика випадкових чисел. Про це красномовно свідчить наступний приклад генератора:

$$\begin{cases}x_i=\dfrac{\widetilde{x}_i}{M},\\ \widetilde{x}_i=(1+\widetilde{x}_{i-1}) \bmod M, \quad i=1,2,\dots\end{cases}$$

Легко бачити, що він має максимальний період, але послідовність $\{\widetilde{x}_i\}_{i\ge 0}$ далека від випадкової. Тому для цього потрібно проводити додаткові дослідження.

Перехід від лінійної функції до квадратичної приводить до *квадратичного конгруентного методу:*

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

який, очевидно, залишає значення максимального періоду без змін: $T_{\max}=M$.

Подальше збільшення максимального періоду можна досягти, зробивши у лінійній змішаній формулі датчика, яка використовується, глибину пам'яті $l>1$:

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

Або взагалі звернутися до конструкції генератора наступного загального вигляду з деякою функцією $g(\cdot,\cdot,\ldots,\cdot)$ від $l\,(l>1)$ попередніх значень $\tilde{x}_{i-1},\tilde{x}_{i-2},\ldots,\tilde{x}_{i-l}$:

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M}, \\ \tilde{x}_i = g(\tilde{x}_{i-1},\tilde{x}_{i-2},\ldots,\tilde{x}_{i-l}), \quad 0 \le \tilde{x}_i \le M-1, \quad i = 1,2,\ldots \end{cases}.$$

Останнє ускладнення структури датчика приводить до того, що можливий найбільший період може досягати значення $T_{\max} = M^l$.

1.3.3. Моделювання дискретних випадкових величин

Скористаємося побудованими датчиками рівномірно розподіленої на відрізку $[0,1)$ випадкової величини для моделювання деяких стохастичних змінних. Почнемо з моделювання дискретної випадкової величини ξ. Нехай вона приймає значення y_i з ймовірностями $p_i = P\{\xi = i\}$, $i = \overline{1,m}$.

Так як $\sum_{i=1}^{m} p_i = 1$, то інтервал $[0,1)$ можна розбити на m підінтервалів:

$$\Delta_1 = [0, p_1), \Delta_2 = [p_1, p_1 + p_2), \ldots, \Delta_i = \left[\sum_{j=1}^{i-1} p_j, \sum_{j=1}^{i} p_j\right), \ldots, \Delta_m = \left[\sum_{j=1}^{m-1} p_j, 1\right)$$

причому довжина інтервалу Δ_i буде дорівнювати p_i $(i = \overline{1,m})$. Це дозволяє запропонувати наступний простий алгоритм моделювання дискретної випадкової величини ξ.

Звертаємося до датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової величини і отримуємо на його виході деяке значення x, яке попаде в один із побудованих підінтервалів. Тоді, якщо $x \in \Delta_i$, то логічно вважати що ξ прийняла значення y_i. Повторивши цю процедуру необхідну кількість раз, отримаємо вибірку потрібного об'єму. (У всіх наведених далі алгоритмах моделювання випадкових величин буде описано лише перший крок.)

Цей алгоритм без проблем переноситься на випадок, коли потрібно моделювати дискретну випадкову величину ξ, яка має злічену множину значень y_i, що приймаються з відповідними ймовірностями p_i, $i = 1,2,3,\ldots$. Наявність рекурентних співвідношень для величин p_i може суттєво спростити процедуру моделювання. Особливо це корисно у нашому випадку злічених множини значень ξ. Для ряду відомих розподілів нескладно виписати потрібні рекуренти.

Приклади. 1. Якщо випадкова величина ξ має *геометричний розподіл* з параметром p $(0 < p < 1)$, то для ймовірностей $p_i = p(1-p)^i$ $(i \ge 0)$ очевидно справедливо $p_{i+1} = p_i(1-p)$, $p_0 = p$, $i \ge 0$.

2. Для *розподілу Пуасона* з параметром $\lambda\,(\lambda > 0)$ у якого $p_i = \dfrac{\lambda^i}{i!}e^{-\lambda}$ $(i \ge 0)$ відповідний рекурент має вигляд $p_{i+1} = p_i \dfrac{\lambda}{i+1}$, $p_0 = e^{-\lambda}$, $i \ge 0$.

Простіший шлях можна запропонувати для моделювання рівномірного дискретного розподілу з $p_i = P\{\xi = i\} = \dfrac{1}{m}$, $i = \overline{1,m}$. Якщо x – вихід датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової величини, то значення потрібної величини отримується по формулі $[1 + mx]$, де $[a]$ – ціла частина від a.

1.3.4. Моделювання неперервних випадкових величин

Перейдемо тепер до аналізу неперервного випадку. Нехай потрібно моделювати неперервну випадкову величину ξ. Позначимо її функцію розподілу через $F(z)$.

Розглянемо випадок коли $F(z)$ – строго монотонна функція. Тоді у ролі реалізації ξ може виступити $F^{-1}(x)$, де x – значення отримане з датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової

величини, а $F^{-1}(\cdot)$ - функція обернена до $F(x)$. Впевнимося у цьому.

Нехай η – рівномірно розподілена на відрізку $[0,1)$ випадкова величина. Проаналізуємо функцію розподілу величини $F^{-1}(\eta)$:

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

Що і треба було довести.

Приклад. Застосуємо останній підхід до моделювання випадкової величини ξ, яка має *показниковий (експоненціальний) розподіл* з параметром $\lambda > 0$:

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

Дійсно, так як обернена функція має вигляд $F^{-1}(y) = -\dfrac{\ln(1-y)}{\lambda}$, то $-\dfrac{\ln(1-\eta)}{\lambda}$ має потрібний показниковий розподіл, де η – величина рівномірно розподілена на інтервалі $[0,1)$. А так як $1-\eta$ теж рівномірно розподілена на інтервалі $[0.1)$. то можна зробити висновок, що величина $-\dfrac{\ln(\eta)}{\lambda}$ має показниковий розподіл із параметром $\lambda > 0$. Тоді у ролі реалізації ξ може виступити $-\dfrac{\ln(x)}{\lambda}$, де x – значення отримане з датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової величини.

Звернемося до моделювання *нормального розподілу* з параметрами m та σ^2. Для цього скористаємося наступним твердженням.

Теорема. Нехай величини η_1, η_2 - незалежні, рівномірно розподілені на інтервалі $[0,1)$. 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 на інтервалі $[0,1)$ величиною. Тоді згідно теореми можна стверджувати, що значення

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

є спостереженнями над незалежними, нормально розподіленими з параметрами m та σ^2 величинами.

У разі необхідності моделювання випадкової величини *рівномірно розподіленої на інтервалі* $[a,b)$, достатньо скористатися очевидним перетворенням виходу x датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової величини: $a + (b-a)x$.

2. Попередня обробка даних

На перших кроках обробки інформації намагаються провести всю роботу пов'язану з отриманням попередніх висновків про змінні, які спостерігаються, підготувати необхідну інформацію для успішного проведення наступних кроків аналізу даних. Саме цим і займається *попередня обробка даних*.

На цьому етапі визначаються основні характеристики вибірки, з неї видаляються аномальні спостереження, перевіряється симетричність розподілу змінної, яка спостерігається, а також здійснюється перевірка вибірки на однорідність, стохастичність, узгодженість з певним законом розподілу і т. п. Для досягнення цих цілей також широко використовують засоби розвідувального аналізу.

2.1. Квантилі та процентні точки розподілу

Ознайомимося з цими двома поняттями, бо вони будуть широко використовуватися у подальшому при розв'язанні задач перевірки гіпотез, побудові довірчих інтервалів і т.п. Визначати будемо окремо для неперервних та дискретних розподілів.

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

Означення 1. *Квантилем рівня q розподілу* (q-квантилем розподілу) *неперервної випадкової величини* ξ називається таке значення u_q, яке визначається з рівняння

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

Означення 2. *Квантилем рівня q розподілу* (q-квантилем розподілу) *дискретної випадкової величини* ξ називається довільне значення u_q з інтервалу $\left(y_{i(q)}, y_{i(q)+1}\right]$, для границь якого справедливо

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

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

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

Означення 3. *Q-процентною точкою розподілу неперервної випадкової величини* ξ називається таке значення w_Q, яке є розв'язком рівняння

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

Означення 4. *Q-процентною точкою розподілу дискретної випадкової величини* ξ називається довільне значення w_Q з інтервалу $\left(y_{i(Q)}, y_{i(Q)+1}\right]$, для границь якого справедливо

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

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

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

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

Ряд характеристик, які базуються на цих поняттях, набули широкого вжитку. Познайомимося з ними.

Приклади.

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. *Процентилі* 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 $\frac{1}{4}$, 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. *Інтердецильна широта* – це інтерквантильна широта рівня $\frac{1}{10}$, а саме $\left(u_{0.9} - u_{0.1}\right)$.

9. *Інтерсектильна широта* – це інтерквантильна широта рівня $\frac{1}{6}$, тобто $\left(u_{\frac{5}{6}} - u_{\frac{1}{6}}\right)$.

Для широко вживаних розподілів складені відповідні таблиці, з яких легко визначити потрібні квантилі та процентні точки [3].

2.2. Характеристики положення центра значень змінної

Ці характеристики є одними з самих широко вживаних в силу їх наглядної інтерпретації. Ознайомимося з ними детальніше.

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

1. *Математичне сподівання* (теоретичне середнє) обчислюється по відомій формулі для $M\xi$. А відповідне вибіркове значення має вигляд:

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

2. *Середнє геометричне* G_ξ визначається для випадкових величин, які з ймовірністю 1 додатні. Згідно з означенням $G_\xi = e^{M\ln(\xi)}$. А його оцінка задається виразом $\hat{G}_\xi(n) = \sqrt[n]{\prod_{i=1}^{n} x_i}$.

3. *Середнє гармонічне* H_ξ вводиться для випадкових величин ξ з позитивними значеннями наступним чином: $H_\xi = \frac{1}{M\left(\frac{1}{\xi}\right)}$. Емпіричне значення цієї характеристики обчислюється згідно з формулою

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

4. *Мода* x_{mod} для неперервної випадкової величини ξ вводиться як точка максимуму функції щільності ξ. Для дискретного розподілу $\{y_i, p_i\}_{i\ge 1}$, з розташованими в порядку зростання значеннями y_i, ця характеристика визначається, як те довільне значення y_k, яке приймається з найбільшою ймовірністю. Очевидно, що мода може бути не єдиною. Тому ця характеристика більш застосовується до унімодальних розподілів. Її вибіркове значення $\hat{x}_{\text{mod}}(n)$ у неперервному випадку визначають по гістограмі, а у дискретному – по полігону частот відповідно.

5. *Медіана* x_{med} – це квантиль рівня 0,5. Її оцінка $\hat{x}_{med}(n)$ обчислюється на основі емпіричної функції розподілу.

2.3. Характеристики розсіювання значень змінної

Характеристики, які розглядаються у цьому розділі, є мірами відхилення спостережень випадкової величини від її характеристики положення центра значень. Вони вказують, як суттєво можуть віддалятися значення випадкової величини від центра зосередження значень. Ознайомимося послідовно з цими характеристиками.

Припустимо, що маємо вибірку об'єму n спостережень $x_1, x_2, \ldots, x_n$ над скалярною змінною ξ. На її основі визначимо потрібні величини.

1. *Дисперсія* σ^2 підраховується згідно формули $\sigma^2 = D\xi = M(\xi - M\xi)^2$. Незміщена оцінка σ^2 має такий вигляд $s^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}(n))^2$. У ряді випадків виявляється більш корисним інше представлення цієї статистики $s^2(n) = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2(n)\right]$.

2. *Стандартне* (середнє квадратичне) *відхилення* σ є коренем квадратним з дисперсії $\sigma = \sqrt{D\xi}$. Автоматично отримуємо його оцінку $s(n) = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2(n)\right]}$.

Зауважимо, що стандартне відхилення для деякої оцінки називають її *стандартною похибкою*. Так, у випадку обробки нормальної вибірки $N(m,\sigma^2)$ об'єму n, стандартна похибка e_ξ оцінки її математичного сподівання $\bar{x}(n)$ визначається таким чином $e_\xi = \frac{\sigma}{\sqrt{n}}$, а відповідне вибіркове значення має вигляд $\hat{e}_\xi(n) = \frac{s(n)}{\sqrt{n}}$.

3. *Коефіцієнт варіації* V_ξ визначається для випадкових величин у яких $M\xi \neq 0$ і підраховується за формулою $V_\xi = \frac{\sqrt{D\xi}}{M\xi}100\%$. Вибіркове значення цієї характеристики обчислюється таким чином:

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

4. *Ймовірнісне відхилення* d_ξ є половиною інтерквартильної широти, тобто $d_\xi = \frac{1}{2}(u_{0.75} - u_{0.25})$, де $u_{0.25}, u_{0.75}$ – нижній та верхній квартилі відповідно. Емпіричне значення цієї характеристики має вигляд $\hat{d}_\xi(n) = \frac{1}{2}(\hat{u}_{0.75} - \hat{u}_{0.25})$. Тут $\hat{u}_{0.25}, \hat{u}_{0.75}$ – вибіркові значення нижнього та верхнього квартилів.

5. *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 $x_{\min}(n), x_{\max}(n)$ – найменше та найбільше значення у цій вибірці.

6. *Інтервал концентрації розподілу* випадкової величини ξ має такий вигляд $\left(M\xi - 3\sqrt{D\xi}, M\xi + 3\sqrt{D\xi}\right)$. А відповідний вибірковий аналог визначається згідно з $\left(\bar{x}(n) - 3s(n), \bar{x}(n) + 3s(n)\right)$.

2.4. Аналіз скошеності та гостроверхості розподілу

Аналізувати розподіл випадкової величини ξ будемо спираючись на отримані спостереження $x_1, x_2, \ldots, x_n$ над нею.

Очевидно, що якщо розподіл ξ симетричний відносно $M\xi$, то всі його непарні центральні моменти $M(\xi - M\xi)^{2k+1}$ будуть дорівнювати нулеві, якщо тільки вони існують. Тому зрозуміло, чому в основі *коефіцієнта асиметрії* – характеристики скошеності розподілу, лежить третій центральний момент:

$$\beta_1 = \frac{M(\xi - M\xi)^3}{\left[M(\xi - M\xi)^2\right]^{\frac{3}{2}}}, \quad D\xi > 0 .$$

Вибіркове значення коефіцієнта асиметрії визначається таким чином:

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

Очевидно, що для симетричних відносно $M\xi$ розподілів $\beta_1 = 0$. У випадку коли $\beta_1 < 0$, то розподіл буде скошеним праворуч, а якщо $\beta_1 > 0$, то розподіл буде скошеним ліворуч. Для нормальних розподілів завжди $\beta_1 = 0$.

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

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

Обчислення емпіричного значення цього коефіцієнта здійснюється за формулою:

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

Для нормального розподілу коефіцієнт ексцесу дорівнює нулеві. Якщо $\beta_2 > 0$, то розподіл, який досліджується, більш гостроверхий ніж нормальний з відповідними параметрами. У випадку коли $\beta_2 < 0$, розподіл буде менш гостроверхим ніж відповідний нормальний.

2.5. Характеристики випадкових векторів

Перейдемо тепер до аналізу q – мірних випадкових векторів ξ. Припустимо, що отримано n спостережень над цим вектором:

$$x_1, x_2, \ldots, x_n, \qquad x_i \in R^q, \ \overline{i,n}.$$

Для цього випадку кількість характеристик, які знайшли своє широке застосування, набагато менша.

Спочатку познайомимося з ***характеристиками положення центра значень***. Фактично, для цього використовують аналоги відповідних характеристик, розглянутих у скалярному випадку. Ознайомимося з ними.

1. *Математичне сподівання* (теоретичне середнє) буде вже представляти собою вектор, а його формула обчислення залишається без змін: $M\xi$. Відповідне вибіркове значення підраховується за тією ж формулою $\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2. *Мода* x_{mod} у неперервному випадку вектора ξ визначається як точка максимуму функції щільності ξ. Для дискретного розподілу це буде те значення, яке набуває ξ з максимальною ймовірністю. Її єдиність гарантується тільки для унімодальних розподілів. Саме для них вона найчастіше і використовується. Її вибіркове значення $\hat{x}_{\text{mod}}(n)$ у неперервному випадку визначають за емпіричною функцією щільності, а у дискретному –за полігоном частот відповідно.

У ролі ***характеристик розсіювання значень*** у векторному випадку виступають коваріаційна матриця та деякі характеристики побудовані на ній. Розглянемо їх.

1. *Коваріаційна матриця* визначається за відомою формулою $\Sigma = M(\xi - M\xi)(\xi - M\xi)^T$. А її оцінку можна підрахувати за формулою:

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

2. *Узагальнена дисперсія* визначається як визначник коваріаційної матриці $\det(\Sigma)$, а її емпіричне значення має вигляд $\det(\hat{\Sigma}(n))$.

3. *Слід коваріаційної матриці*, очевидно, підраховується таким чином $tr(\Sigma)$. Вибіркове значення цієї характеристики, в свою чергу, обчислюється згідно з виразом $tr(\hat{\Sigma}(n))$.

2.6. Перевірка стохастичності вибірки

Перш ніж проводити обробку досліджуємої вибірки має сенс впевнитися, що вона є випадковою, а не знаходиться під впливом деякого систематичного зміщення, наприклад: монотонного або періодичного. Для перевірки стохастичності вибірки були запропоновані наступні критерії:

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

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

Критерій серій на базі медіани вибірки. Спочатку визначається оцінка медіани $\hat{x}_{med}$. Потім під кожним членом вибірки $x_1, x_2, \ldots, x_n$, який більше $\hat{x}_{med}$ ставиться плюс, а який менше $\hat{x}_{med}$ ставиться мінус. Виміри які дорівнюють $\hat{x}_{med}$ до уваги не приймаються. Для отриманої таким чином послідовності плюсів та мінусів обчислюємо дві статистики: загальну кількість серій $\nu(n)$ у ній та кількість членів у найдовшій серії $\tau(n)$. Під *серією* мається на увазі послідовність підряд розташованих однакових символів плюс чи мінус. Зрозуміло, що вибірка буде мати стохастичну природу, якщо довжина найдовшої серії $\tau(n)$ не занадто довга, загальна кількість серій не занадто мала. Тоді спираючись на статистики $\nu(n)$ та $\tau(n)$, можна записати область прийняття нашої гіпотези:

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

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

Критерій зростаючих та спадаючих серій. Цей критерій чутливий до наявності у вибірці не тільки монотонних, але й циклічних зміщень середнього. Спочатку у вибірці $x_1, x_2, \ldots, x_n$ замінюємо підряд розташовані однакові виміри одним їх представником. Потім під кожним членом таким чином трансформованої вибірки ставимо символ плюс, якщо його наступний член з вибірки строго більше поточного. І ставимо символ мінус, якщо його наступний член з вибірки строго менше поточного. Далі на базі таким чином утвореної послідовності плюсів та мінусів визначаємо статистики $\nu(n)$ та $\tau(n)$ абсолютно аналогічно, як це робилося у попередньому критерії. Далі використовується та ж сама ідея для побудови області прийняття нашої гіпотези про стохастичність вибірки. Вона буде мати ідентичний вигляд:

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

де $\nu_\beta(n), \tau_\beta(n)$ – квантилі рівня β статистик $\nu(n)$ та $\tau(n)$ відповідно. Гіпотеза сприймається тільки у випадку справедливості обох нерівностей. Зроблене для попереднього критерію зауваження відносно можливих значень для α залишається у силі.

Критерій квадратів послідовних різниць (критерій Аббе). Даний критерій використовується при роботі з нормальними вибірками, бо він виявляється більш потужним ніж попередні на цьому класі вибірок. У ролі альтернативи при перевірці нашої гіпотези тут може виступати наявність систематичного зміщення у вибірці.

На основі вибірки підрахуємо наступну статистику:

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

Область прийняття гіпотези для цього критерію має вигляд:

$$\gamma(n) > \gamma_\alpha(n),$$

де $\gamma_\alpha(n)$ визначається по відповідній таблиці з роботи [3], якщо $n \le 60$, в протилежному випадку потрібно скористатися формулою:

$$\gamma_\alpha(n) = 1 + \frac{u_\alpha}{\sqrt{n + 0{,}5\left(1 + u_\alpha^2\right)}}.$$

Тут u_α – квантиль рівня α нормального розподілу з параметрами 0 та 1.

2.7. Рангові критерії однорідності

Розглянемо випадкові величини $\xi_1, \xi_2, \ldots, \xi_k$ з функціями розподілу $F_1(x), F_2(x), \ldots, F_k(x)$. Нехай для кожної змінної ξ_i отримані незалежні спостереження

$$x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}, \quad i = \overline{1,k}.$$

На їх основі сформуємо *об'єднану вибірку* $v_1, v_2, \ldots, v_n$ об'єму $n = \sum_{i=1}^{k} n_i$. Для спрощення будемо вважати, що всі виміри v_i $(i = \overline{1,n})$ різні. (У подальшому випадок наявності нерозрізнимих спостережень буде розглянуто окремо.) Розташувавши ці значення у порядку зростання, отримаємо відповідний *варіаційний ряд* $v_{(1)}, v_{(2)}, \ldots, v_{(n)}$. Члени останнього ряду називають *порядковими статистиками*.

Означення. *Рангом* спостереження v_i $\left(i = \overline{1,n}\right)$ називається його порядковий номер у побудованому варіаційному ряді $v_{(1)}, v_{(2)}, \ldots, v_{(n)}$.

Позначимо $R_{i,n}$ – ранг спостереження v_i $\left(i = \overline{1,n}\right)$. Рангові критерії однорідності базуються на використанні у своїх статистиках саме цих значень рангів $R_{i,n}$ $\left(i = \overline{1,n}\right)$. Будемо розглядати *лінійні рангові критерії* з наступними статистиками:

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

Тобто підсумовування у статистиці K_i відбувається тільки по спостереженням над i-тою змінною ξ_i. (Величини $\varphi(R_{j,n})$ називають *мітками*.) Привабливим є те, що розподіли далі розглянутих такого роду статистик, прямують до нормальних законів з деякими параметрами при зростанні об'ємів вибірок.

Нехай потрібно перевірити гіпотезу:

$$H_0: \quad F_1(x) = F_2(x) = \ldots = F_k(x), \forall x,$$

з деяким рівнем значимості $\alpha \quad (0 < \alpha < 1)$. Тобто необхідно впевнитися, що величини $\xi_1, \xi_2, \ldots, \xi_k$ однаково розподілені. Спочатку буде розглянуто ситуацію $k = 2$, а потім перенесено результати на загальний випадок.

2.7.1. Випадок двох вибірок

Нехай спостерігаються випадкові величини ξ_1 та ξ_2 з функціями розподілу $F_1(x)$ та $F_2(x)$ відповідно. Перевіряємо гіпотезу

$$H_0: \quad F_1(x) = F_2(x), \forall x,$$

з деяким рівнем значимості $\alpha \quad (0 < \alpha < 1)$.

Як альтернативну гіпотезу розглянемо такі її варіанти:

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

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

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

Тобто розподіли відрізняються своїми характеристиками положення центра значень. А саме спостерігається зміщення розподілу $F_2(x)$ по відношенню до $F_1(x)$ відповідно: довільне, вліво або вправо.

Для перевірки нульової гіпотези було запропоновано ряд критеріїв. Зазначимо, що в розглянутих нижче статистиках підсумовування здійснюється тільки за спостереженнями над першою змінною ξ_1. Ознайомимося з ними та їх характеристиками.

Критерій нормальних міток (Фішера). Його статистика має такий вигляд:

$$C = \sum_{i=1}^{n_1} M(R_{i,n}, n),$$

де $M(m,n)$ – математичне сподівання m-тої порядкової статистики вибірки довжини $n = n_1 + n_2$ нормально розподіленої величини з параметрами 0 та 1. Відповідна апроксимаційна формула для обчислення значень $M(m,n)$ наведена у додатку 2. Статистика C має наступні характеристики при справедливості нульової гіпотези:

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

Критерій Ван дер Вардена. В його основі лежить статистика, яка задається наступним виразом:

$$V = \sum_{i=1}^{n_1} \Phi^{-1}\left(\frac{R_{i,n}}{n+1}\right),$$

де $\Phi^{-1}(x)$ – функція обернена до функції розподілу нормального закону з параметрами 0 та 1. При обчисленні значень $\Phi^{-1}(x)$ можна скористатися наближеною формулою, яка міститься у додатку 2.

Якщо справедлива нульова гіпотеза, то:

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

Критерій Вілкоксона. Цей критерій базується на використанні статистики, яка має дуже простий вигляд:

$$S = \sum_{i=1}^{n_1} R_{i,n}.$$

Коли нульова гіпотеза вірна, то для статистики S справедливо:

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

Процедура використання статистик C, V та S для перевірки гіпотези H_0 однакова. Тому наведемо один спільний алгоритм для них, скориставшись раніше згаданою асимптотичною властивістю розподілів цих статистик.

Позначимо через U статистику одного з вищенаведених критеріїв (тобто C, V або S). Підрахуємо її стандартизоване значення:

$$\overline{U} = \frac{U - MU}{\sqrt{DU}},$$

де MU, DU – математичне сподівання та дисперсія статистики U відповідно. В результаті область прийняття гіпотези H_0 буде залежати від вигляду альтернативної гіпотези:

$|\overline{U}| < u_{1-\frac{\alpha}{2}}$, якщо альтернатива H_{11}.

$\overline{U} < u_{1-\alpha}$, якщо альтернатива H_{12},

$\overline{U} > u_{\alpha}$, якщо альтернатива H_{13},

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

При великих об'ємах вибірок порівняння розглянутих критеріїв дозволяє розташувати їх наступним чином по мірі спадання потужності: критерій нормальних міток, критерій Ван дер Вардена, критерій Вілкоксона.

2.7.2. Загальний випадок

Перенесемо результати, отримані у попередньому розділі, на випадок аналізу більше двох вибірок. Тобто вважаємо, що спостерігаємо випадкові величини $\xi_1, \xi_2, \ldots, \xi_k$ з функціями розподілу $F_1(x), F_2(x), \ldots, F_k(x)$. Потрібно перевірити гіпотезу

$$H_0: \quad F_1(x) = F_2(x) = \ldots = F_k(x), \forall x,$$

з деяким рівнем значимості $\alpha \quad (0 < \alpha < 1)$.

Для кожної змінної ξ_i маємо у розпорядженні її незалежні виміри $x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}, \quad i = \overline{1,k}$. На їх базі будуємо об'єднану вибірку $v_1, v_2, \ldots, v_n$ об'єму $n = \sum_{i=1}^{k} n_i$, потім відповідний варіаційний ряд $v_{(1)}, v_{(2)}, \ldots, v_{(n)}$.

Продемонструємо процедуру перевірки гіпотези H_0 у загальному випадку. З цією метою для кожної змінної ξ_i, скориставшись одним із лінійних рангових критеріїв, підраховуємо статистику:

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

причому підсумовування у статистиці K_i проводимо тільки по спостереженням над i-тою змінною ξ_i $(i = \overline{1,k})$. Якщо використовувати розглянуті у попередньому розділі критерії, то це будуть фактично статистики

$$C_i = \sum_{j=N_i-n_i+1}^{N_i} M(R_{j,n}, n), \; V_i = \sum_{j=N_i-n_i+1}^{N_i} \Phi^{-1}\left(\frac{R_{j,n}}{n+1}\right) \text{ або } S_i = \sum_{j=N_i-n_i+1}^{N_i} R_{j,n}$$

відповідно $(i = \overline{1,k})$. Далі знаходимо їх стандартизовані значення

$$\overline{K}_i = \frac{K_i - MK_i}{\sqrt{DK_i}},$$

де MK_i, DK_i – математичне сподівання та дисперсія статистики K_i відповідно $\left(i = \overline{1,k}\right)$.

А так як усі статистики K_i асимптотично нормальні, то для перевірки H_0 скористаємося статистикою $X^2 = \sum_{i=1}^{k} \overline{K}_i^2$. У підсумку, область прийняття нульової гіпотези виглядає

$$X^2 < \chi_{\alpha}^2(k-1),$$

де $\chi_{\alpha}^2(k-1)$ – $100\alpha\,\%$ процентна точка χ^2-розподілу з $(k-1)$ степенями свободи.

2.8. Перевірка симетрії розподілу ранговими критеріями

Припустимо, що отримано ряд незалежних спостережень:

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

над випадковою величиною ξ з функцією розподілу $F(x)$. Розглянемо задачу перевірки симетричності розподілу цієї величини відносно деякої точки x_0. За останню дуже часто вибирають одну з характеристик положення центра значень змінної. Саму гіпотезу формально можна записати наступним чином:

$$H_0: \quad F(x_0 + x) = 1 - F(x_0 - x + 0), \quad \forall x.$$

Перевірку проводимо з деяким рівнем значимості $\alpha \quad (0 < \alpha < 1)$. Якщо ж маємо справу з неперервним розподілом з функцією щільності $p(x)$, то нульова гіпотеза набуває вигляду:

$$H_0: \quad p(x_0 + x) = p(x_0 - x), \quad \forall x.$$

Розглянемо можливості застосування рангових критеріїв до перевірки гіпотези H_0. Для цього на основі вибірки $x_1, x_2, \ldots, x_n$ побудуємо послідовність:

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

де $z_i = |x_i - x_0|$, $i = \overline{1,n}$. А далі формуємо відповідний варіаційний ряд $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$.

Означення. *Абсолютним рангом виміру* x_i називається порядковий номер значення $z_i = |x_i - x_0|$ у варіаційному ряді $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$. Будемо використовувати $R^+_{i,n}$ для його позначення $\left(i = \overline{1,n}\right)$.

Розіб'ємо вибірку $x_1, x_2, \ldots, x_n$ на дві підвибірки, а саме на ті виміри які більші x_0 та всі інші. Введемо позначення для індексів спостережень з першої підвибірки $I^+ = \left\{i : x_i > x_0, i = \overline{1,n}\right\}$. Тепер можна застосувати один з рангових критеріїв перевірки однорідності для випадку двох вибірок до утворених таким чином двох підвибірок. Зауважимо, що скрізь у відповідних статистиках підсумовування будемо здійснювати тільки за спостереженнями з першої підвибірки, тобто за множиною I^+.

Аналог критерію нормальних міток. Його статистика має наступний вигляд:

$$C^+ = \sum_{i \in I^+} M^-(R^+_{i,n}, n),$$

де $M^-(m,n)$ – математичне сподівання m-тої порядкової статистики вибірки довжини n модуля нормально розподіленої величини з параметрами 0 та 1. При справедливості нульової гіпотези статистика C^+ має такі характеристики:

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

Аналог критерію Ван дер Вардена. Він базується на використанні статистики виду:

$$V^+ = \sum_{i \in I^+} \Phi^{-1}\left(\frac{1}{2} + \frac{R^+_{i,n}}{2(n+1)}\right),$$

де $\Phi^{-1}(x)$ – обернена функція до функції розподілу нормального закону з параметрами 0 та 1. Якщо справедлива H_0, то

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

Аналог критерію Вілкоксона. Цей критерій використовує статистику вигляду:

$$S^+ = \sum_{i \in I^+} R^+_{i,n}.$$

У випадку справедливості нульової гіпотези можна стверджувати, що

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

Скориставшись асимптотичною нормальністю розподілів статистик C^+, V^+ та S^+, наведемо спільний алгоритм перевірки гіпотези H_0. Нехай U^+ – статистика одного з вищенаведених критеріїв (тобто C^+, V^+ або S^+). Визначимо її стандартизоване значення:

$$\overline{U}^+ = \frac{U^+ - MU^+}{\sqrt{DU^+}},$$

де MU^+, DU^+ – математичне сподівання та дисперсія статистики U^+ відповідно. В результаті область прийняття гіпотези H_0 буде мати вигляд:

$$\left|\overline{U}\right| < u_{1-\frac{\alpha}{2}},$$

де u_α – квантиль рівня α нормального закону з параметрами 0 та 1.

2.9. Визначення рангів у випадку наявності нерозрізнимих значень

До цього моменту при використанні рангових критеріїв припускалося, що усі спостереження набувають різних значень. Звернемося тепер до загального випадку, коли допустима наявність нерозрізнимих значень. Причиною їх появи може бути групування даних, дискретна природа величин, які спостерігаються і т. п. Наявність груп рівних значень спостережень вимагає внесення корекції у процедуру присвоєння рангів цим вимірам.

Нехай $v_1, v_2, \ldots, v_n$ – об'єднана вибірка, побудована на основі спостережень над змінними, які досліджуються. Для нерозрізнимих

значень можливі дві ситуації, а саме вони можуть бути з підвибірки спостережень над однією змінною або з підвибірок спостережень над різними змінними.

У першому випадку фактично ніякої проблеми немає. Цим вимірам з групи нерозрізнимих значень призначають ранги у довільному порядку з множини номерів, які припали на цю групу спостережень.

Другий випадок потребує більш детальнішого розгляду. Ранги у цій ситуації можна присвоювати, використовуючи різні методи. Розглянемо деякі з відомих підходів.

Метод випадкового рангу. Він полягає у тому, що рівним значенням, ранги призначають випадковим чином, розігруючи рівноймовірнісно ранги, які припали на цю групу нерозрізнимих значень. Все інше залишається без змін у процедурі використання критеріїв, які були розглянуті раніше.

Метод середньої мітки. При цьому підході всім рівним спостереженням присвоюють середнє значення мітки підраховане за множиною рангів, яка відповідає цій групі нерозрізнимих вимірів. А в алгоритмі застосування рангового критерію потребує корекції лише значення дисперсії статистики, яка використовується. Наприклад, для статистики C критерію нормальних міток відповідне значення дисперсії набуде вигляду:

$$DC = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^{g} \tau_i \overline{M}_i^2 ,$$

а для критерію Ван дер Вардена в свою чергу

$$DV = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^{g} \tau_i \left(\overline{\Phi}_i^{-1} \right)^2 ,$$

де g – кількість груп нерозрізнимих спостережень, τ_i – кількість значень в i - тій групі, $\overline{M}_i$ та $\overline{\Phi}_i^{-1}$ – середні значення міток для i - тої групи рівних вимірів для критерію нормальних міток та Ван дер Вардена відповідно ($i = \overline{1, g}$).

Зауваження. Потужність вищерозглянутих критеріїв при використанні методу середньої мітки буде більшою ніж при застосуванні методу випадкового рангу.

2.10. Видалення аномальних спостережень

Перш за все, з'ясуємо, які спостереження потрібно вважати аномальними, або як їх ще називають *викидами*. До *аномальних спостережень* будемо відносити ті виміри, значення яких не узгоджуються з розподілом більшості отриманих спостережень. Поява їх може пояснюватися різними причинами: порушення умов проведення експерименту, похибки приладів, які використовуються, збій у роботі обладнання, стихійні лиха, інші непередбачувані причини тощо.

Так як найбільш розробленою виявилася теорія для нормальних вибірок, то саме їх і будемо розглядати.

2.10.1. Обробка скалярних вимірів

У цьому випадку для виявлення викидів запропоновано ряд методів, а саме: критерій Граббса, критерій Томпсона, критерій Тітьєна-Мура, а також графічні методи на основі пробіт-графіку, ймовірнісного графіку, зображення "стебло-листок", зображення "скринька з вусами", які будуть розглянуті у розвідувальному аналізі.

Ще раз зазначимо, що критерії, які розглядаються нижче, розроблені для нормальних вибірок. Нехай саме така вибірка $x_1, x_2, \ldots, x_n$ отримана для обробки. Будемо перевіряти гіпотезу H_0, що найбільш підозрілий на аномальність вимір не є викидом, з рівнем значимості α $(0 < \alpha < 1)$.

Критерій Граббса. Спочатку за вибіркою підрахуємо такі статистики:

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

Далі будуємо послідовність:

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

де $z_i = |x_i - \bar{x}(n)|$, $i = \overline{1, n}$. А потім відповідний варіаційний ряд:

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

Тут $z_{(j)} = |x_{i(j)} - \bar{x}(n)|$, $j = \overline{1.n}$. Логічно тепер підозрювати на аномальність спостереження, яке відповідає останньому члену варіа-

ційного ряду $z_{(n)}$, а саме $x_{i(n)}$. Перевіримо його на аномальність. Для цього обчислимо наступну статистику:

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

Тоді логічно за область прийняття нашої гіпотези, що $x_{i(n)}$ не є викидом, взяти таку:

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

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

Якщо спостереження $x_{i(n)}$ виявилося аномальним, то його видаляють з вибірки і всю процедуру повторюють, але вже з цією скороченою вибіркою. Все це повторюється до того часу доки на деякому кроці найбільш підозрілий на аномальність вимір виявиться не викидом. Після цього роботу завершують.

Критерій Томпсона. Цей критерій є фактично деякою модифікацією попереднього. Перші кроки, включаючи обчислення статистики $T(n)=\frac{x_{i(n)}-\bar{x}(n)}{s(n)}$ залишилися без змін. Далі будується така статистика:

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

І область прийняття нашої гіпотези набуде вигляду:

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

де $t_{\frac{\alpha}{2}}(n-2)$ - $100\frac{\alpha}{2}\%$ точка t-розподілу Стьюдента з $(n-2)$ степенями свободи. (Тобто вдалося перейти до використання t-розподілу Стьюдента, що більш зручніше.)

Надалі, якщо $x_{i(n)}$ виявився аномальним, то його видаляють з вибірки і всю процедуру повторюють по тій же схемі що й в критерії Граббса.

Критерій Тітьєна-Мура. На відміну від попередніх критеріїв тут є можливість перевірити на аномальність одразу декілька найбільш

підозрілих спостережень. Знову підраховуємо $\bar{x}(n)=\frac{1}{n}\sum_{i=1}^{n}x_i$, потім будуємо послідовність:

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

де $z_i=|x_i-\bar{x}(n)|$, $i=\overline{1,n}$. Далі формуємо відповідний варіаційний ряд:

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

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

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

Тепер обчислимо статистику:

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

де $\bar{z}(m)=\frac{1}{m}\sum_{i=1}^{m}z_{(i)}$. Тоді область прийняття нашої гіпотези, що спостереження $x_{i(n-k+1)}, x_{i(n-k+2)}, \ldots, x_{i(n)}$ не є викидами, при перевірці її з рівнем значимості α $(0<\alpha<1)$ буде мати наступний вигляд:

$$E(n,k)\geq E_{1-\alpha}(n,k),$$

де $E_{1-\alpha}(n,k)$ - $100(1-\alpha)\%$ точка 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 з скороченою вибіркою, інакше алгоритм завершує свою роботу.

2.10.2. Випадок векторних значень

Припустимо тепер, що спостереження є векторними величинами:

$$x_1, x_2, \ldots, x_n, \qquad x_i\in R^q, \ i=\overline{1,n}.$$

Будемо перевіряти гіпотезу H_0, що найбільш підозрілий на аномальність вектор вимірів не є викидом, з рівнем значимості α $(0<\alpha<1)$. Пропонується скористатися *критерієм на базі F - статистики.*

Спочатку обчислюємо наступні величини:

$$\bar{x}_i = \frac{1}{n-1}\sum_{j\neq i} x_j, \quad \hat{\Sigma}_i = \frac{1}{n-2}\sum_{j\neq i}(x_j-\bar{x}_i)(x_j-\bar{x}_i)^T, \quad i=\overline{1,n}.$$

Потім підраховуємо вибіркові відстані Махаланобіса:

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

Визначаємо значення наступних статистик:

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

Знаходимо індекс найбільш підозрілого на аномальність виміру

$$i_0 = \arg\max_i F_i.$$

Тоді гіпотеза H_0 буде прийматися, якщо справедлива наступна нерівність:

$$F_{i_0} < F_\alpha(q, n-1-q),$$

де $F_\alpha(q,n-1-q)$ – $100\alpha\%$ точка F – розподілу з q та $n-1-q$ степенями свободи.

Якщо вектор спостережень x_{i_0} виявився аномальним, то його видаляють з вибірки і всю процедуру повторюють до того часу доки на деякому кроці найбільш підозрілий на аномальність вимір виявиться не викидом. Після цього процедуру завершують.

3. Розвідувальний аналіз

Розвідувальний аналіз – це один із етапів попередньої обробки даних, який дозволяє провести візуально експрес-аналіз отриманих даних на основі їх представлення у вигляді графіків, схем або таблиць. Результати цього аналізу слугуватимуть відправною точкою для планування подальшої поглибленої обробки інформації.

У випадку спостережень над однією змінною у розвідувальному аналізі можна використовувати: пробіт-графік (probit plot), ймовірнісний графік (probability plot), висячі гістобари (hanging histobars), підвішену коренеграму (suspended rootogram), зображення "стеблолисток" (stem-and-leaf plot), зображення "скринька з вусами" (box-and-whisker plot) та його модифікації і т.д. Далі будуть розглянуті послідовно можливості кожного з цих засобів.

3.1. Сімейства розподілів типу зсув-масштабу

Перші два графічних представлення даних можуть використовуватися для змінних функції розподілу, що належать до сімейств типу зсув-масштабу.

Означення. Сімейство розподілів F називається *сімейством типу зсув-масштабу*, якщо існує така *базова* функція розподілу $F_0(\cdot)\in F$, що для будь-якої функції розподілу $F(\cdot)$ з цього сімейства існують a та b $(b>0)$ такі, що її можна представити наступним чином:

$$F(x) = F_0\left(\frac{x-a}{b}\right).$$

Наведемо приклади сімейств типу зсув-масштабу.

1. Сімейство нормальних розподілів.

Дійсно довільну функцію розподілу $F(x)$ нормально розподіленої величини $\xi \sim N(m,\sigma^2)$ можна представити у вигляді $F(x) = \Phi\left(\frac{x-m}{\sigma}\right)$, де $\Phi(x)$ – функція розподілу нормально розподіленої величини з параметрами 0 та 1. Для цього сімейства розподілів: $\Phi(\cdot)$ – базова функція, $a=m$, $b=\sigma$.

2. Сімейство показникових (експоненціальних) розподілів.

Для функції показникового розподілу $F(x)$ з параметром λ $(\lambda>0)$ справедливо $F(x)=\Phi_1(\lambda x)$, де $\Phi_1(x)$ – функція експоненціального розподілу з параметром 1. Тобто роль базової функції

тут відіграє функція $\Phi_1(\cdot)$, а потрібні константи визначаються згідно з $a=0,\ b=\lambda^{-1}$.

3.2. Пробіт- та ймовірнісний графіки

Далі будемо розглядати обробку вибірки об'єму n спостережень $x_1, x_2, \ldots, x_n$ над скалярною змінною ξ з функцією розподілу $F_\xi(x)$.

Пробіт-графік. Познайомимося з його побудовою. По вибірці $x_1, x_2, \ldots, x_n$ обчислюємо емпіричну функцію розподілу $\hat{F}_\xi(x)$. Пробіт-графік деякого сімейства розподілів F типу зсув-масштабу з базовою функцією $F_0(\cdot)$ задається таким чином:

$$y = F_0^{-1}(\hat{F}_\xi(x)).$$

Подивимося, який повинен мати вигляд побудований пробіт-графік у випадку коли функція розподілу випадкової величини ξ, яка спостерігається, належить цьому сімейству розподілів F типу зсув-масштабу. Тоді існують a та b $(b>0)$ такі, що $\hat{F}_\xi(x) \approx F_0\left(\frac{x-a}{b}\right)$. А сам пробіт-графік буде мати наступний вид:

$$y = F_0^{-1}(\hat{F}_\xi(x)) \approx F_0^{-1}\left(F_0\left(\frac{x-a}{b}\right)\right) = \frac{x}{b} - \frac{a}{b}.$$

Це дозволяє використовувати цей графік для візуального розв'язку наступних задач:

1) Перевірки гіпотези $H_0 : F_\xi(\cdot) \in F$.

   У випадку справедливості цієї гіпотези пробіт-графік буде уявляти собою приблизно деяку пряму, в протилежному випадку гіпотезу відхиляють.

2) Виявлення наявності аномальних спостережень у вибірці.

   Про присутність викидів у вибірці буде говорити наявність деяких точок графіку, які розташовані осторонь основної маси точок графіку.

Ймовірнісний графік. Нехай $\hat{F}_\xi(x)$ – емпірична функція розподілу, яка обчислена по вибірці спостережень $x_1, x_2, \ldots, x_n$ над випадковою величиною ξ. Ймовірнісний графік – це графік функції $y = \hat{F}_\xi(x)$, побудований на спеціальному *ймовірнісному папері* деякого *сімейства розподілів* F типу зсув-масштабу з базовою функцією $F_0(\cdot)$. Останній відрізняється від звичайного паперу зміненим масштабом по осі y. З цією метою на такому папері смугу $\{(x,y): 0 \le y \le 1\}$ трансформують таким чином: $(x,y) \mapsto (x, F_0^{-1}(y))$. Можливості та методика використання ймовірнісного графіку точно такі як і у пробіт-графіку. Якщо F – сімейство нормальних розподілів, то цей графік називають *нормальним ймовірнісним графіком*, а відповідний папір – *нормальним ймовірнісним папером*.

3.3. Візуальні методи перевірки нормальності

Висячі гістобари. Це графічне зображення будується таким чином. Спочатку по вибірці $x_1, x_2, \ldots, x_n$ визначають вибіркові значення математичного сподівання $\hat{m}$ та дисперсії $\hat{\sigma}^2$. Потім будується графік щільності нормального розподілу $N(\hat{m}, \hat{\sigma}^2)$. Далі у центрі кожного інтервалу групування даних до цієї кривої підвішують гістобару, довжина якої пропорційна відносній частоті попадання вимірів у цей інтервал групування.

Висячі гістобари використовують для візуальної перевірки гіпотези нормальності розподілу випадкової величини, яка спостерігається. Гіпотезу приймають, якщо основи гістобар незначно відхиляються від осі абсцис. В протилежному випадку її відхиляють.

Підвішена коренеграма. Вона представляє собою послідовність прямокутників, побудованих у центрах інтервалів групування даних вибірки $x_1, x_2, \ldots, x_n$, довжини яких пропорційні різниці $\sqrt{\nu} - \sqrt{\nu_0}$, де ν, ν_0 – частоти попадання у інтервал групування емпірична та підрахована згідно нормального розподілу $N(\hat{m}, \hat{\sigma}^2)$ відповідно (тут $\hat{m}$ та $\hat{\sigma}^2$ – вибіркові значення середнього та дисперсії).

Це графічне представлення можна використовувати для візуальної перевірки гіпотези нормальності розподілу випадкової величини, яка спостерігається. Остання вважається нормально розподіленою, якщо побудовані прямокутники незначно відхиляються від осі абсцис, інакше вона відхиляється.

3.4. Інші графічні методи

Існує ще ряд візуальних підходів експрес-аналізу даних, які надають можливість отримати своєрідну корисну інформацію за спостереженнями. Ознайомимося з деякими з цих графічних методів безпосередньо.

Зображення "стебло-листок". Це зображення дозволяє представити отримані дані наглядним чином. Розглянемо його на конкретному прикладі.

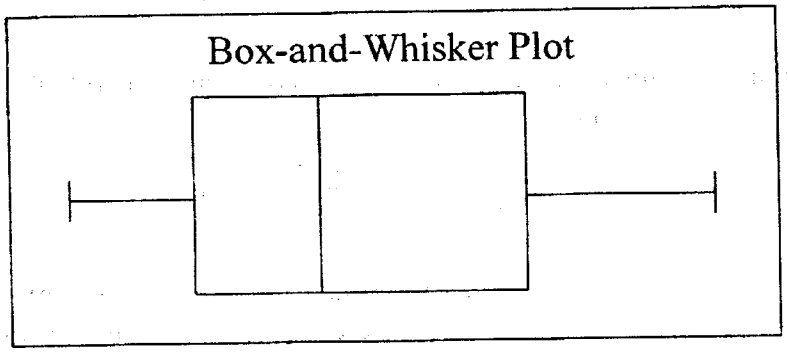
```
Stem-and-Leaf Plot for Variable1: unit = 100      1|2  represents 1200
      LO |18, 19, 21, 21
 7    2F |455
18    2S |66666777777
30    2° |888888999999
46    3* |0000000001111111
66    3T |22222223333333333333
84    3F |444444444555555555
106   3S |66666666667777777777
73    3° |88888888888999999999999999
47    4* |0000000000111111
32    4T |22333
27    4F |444444455555555
12    4S |6666777
 5    4° |89
 3    5* |0
      HI |61, 73
```

У першому рядку вказано, що це зображення побудовано для змінної Variable1, використовуючи масштабний множник 100. Ліві-

ше вертикальної риски вказується ведуча цифра поточного виміру з одним із фіксованих символів, а правіше вертикальної риски його наступна цифра. Врахувавши масштабний множник 100, одразу отримаємо значення поточного спостереження. Цифра, яка стоїть у першому стовпчику вказує кількість відображених спостережень у поточному рядку плюс у всіх рядках до найближчого краю зображення. У самих крайніх рядках, які починаються з абревіатур LO або HI, можуть вказуватися виміри підозрілі на аномальність.

Зображення "стебло-листок" дозволяє візуально з'ясувати загальний вигляд розподілу даних, інтервал їх концентрації, симетричність розподілу, наявність аномальних вимірів.

Зображення "скринька з вусами". Воно має у загальному випадку такий нижченаведений вигляд:



Проекція середньої вертикальної лінії скриньки на вісь абсцис дає нам значення медіани, лівої границі скриньки – нижнього квартилю, правої границі скриньки – верхнього квартилю. Проекції лівого кінця лівого вуса та правого кінця правого вуса відповідно дають нам найменше найбільше значення у вибірці. При наявності викидів на зображенні вони з'являються у вигляді окремих точок, відображених лівіше та правіше кінців вищевказаних ліній.

Додаток 1. Нормальний закон та пов'язані з ним розподіли

Випадкові величини, які мають нормальний розподіл або розподіли похідні від нього відіграють суттєву роль при аналізі даних.

Коротко дамо огляд найбільш широко вживаним розподілам, котрі будуються на базі гауссовського закону.

Означення 1. Випадкова величина χ_n^2 має *χ^2-розподіл з n степенями свободи*, якщо її функція щільності задається таким чином:

$$p(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}x^{\frac{n}{2}-1}e^{\frac{x}{2}}, & \text{якщо } x>0,\\ 0, & \text{якщо } x\le 0,\end{cases}$$

де $\Gamma(n)$ – значення гамма-функції у точці n, яке в загальному випадку визначається таким чином:

$$\Gamma(x)=\int_0^\infty e^{-y}y^{x-1}dy, \quad x>0.$$

Для підрахування потрібних значень будуть корисні формули:

$$\Gamma(m)=(m-1)!,$$

$$\Gamma\left(m+\frac{1}{2}\right)=\frac{1\cdot 3\cdot 5\cdots(2m-1)}{2^m}\sqrt{\pi},$$

де m – ціле позитивне число.

Більш наглядно цей розподіл визначається таким чином. Нехай випадкові величини $\xi_1,\xi_2,\ldots,\xi_n$ – незалежні, нормально розподілені з параметрами 0 та 1. Тоді величина $\chi_n^2=\xi_1^2+\xi_2^2+\ldots+\xi_n^2$ має χ^2-розподіл з n степенями свободи.

Характеристики цього розподілу:

$$M\chi_n^2=n, \quad D\chi_n^2=2n.$$

Означення 2. Випадкова величина t_n має *t-розподіл (Стьюдента) з n степенями свободи*, якщо її функція щільності дорівнює

$$p(x)=\frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\dfrac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Цей розподіл можна визначити іншим шляхом. Нехай незалежні випадкові величини ξ_0,χ_n^2 мають стандартний нормальний розподіл та χ^2-розподіл з n степенями свободи відповідно. Тоді величина

$$t_n=\frac{\xi_0}{\sqrt{\dfrac{\chi_n^2}{n}}}$$

має t-розподіл з n степенями свободи.

Можна впевнитися, що

$$Mt_n=0,$$

$$Dt_n=\frac{n}{n-2}, \text{ якщо } n>2.$$

Означення 3. Випадкова величина $F_{m,n}$ має *F-розподіл (Фішера-Снедекора) з m та n степенями свободи*, якщо її функція щільності визначається

$$p(x)=\begin{cases}\dfrac{\Gamma\left(\dfrac{m+n}{2}\right)m^{\frac{m}{2}}n^{\frac{n}{2}}}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)}x^{\frac{m}{2}-1}\left(n+mx\right)^{-\frac{m+n}{2}}, & \text{якщо } x>0.\\ 0, & \text{якщо } x\le 0.\end{cases}$$

Більш прозоро цей розподіл можна визначити таким шляхом. Нехай незалежні випадкові величини χ_m^2,χ_n^2 мають χ^2-розподіл з m та n степенями свободи відповідно. Тоді величина

$$F_{m,n}=\frac{\dfrac{\chi_m^2}{m}}{\dfrac{\chi_n^2}{n}}$$

буде мати F-розподіл з m та n степенями свободи.

Для цього розподілу можна підрахувати

$$MF_{m,n} = \frac{n}{n-2}, \text{ якщо } n > 2$$

$$DF_{m,n} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \text{ якщо } n > 4.$$

Додаток 2. Характеристики порядкових статистик

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

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

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

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 є можливість скористатися більш грубим наближенням, обмежившись тільки першим доданком $Mx_{(m)} \approx \Phi^{-1}(\alpha_m)$.

Залишилося з'ясувати питання обчислення значень $\Phi^{-1}(\alpha), \alpha \in (0,1)$. Використаємо відповідну апроксимаційну формулу.

Коли $\alpha \in [0.5, 1)$, то

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

де $\tau = \sqrt{-2\ln(1-\alpha)}$, $a_0 = 2.515517, a_1 = 0.802853, a_2 = 0.010328,$

$b_1 = 1.432788, b_2 = 0.189269, b_3 = 0.001308$.

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

Список літератури

1. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1983.
2. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ. – М.: Мир, 1982.
3. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М.: Наука, 1983.
4. *Кнут Д.Э.* Искусство программирования: Т.2. Получисленные алгоритмы, 3-е изд. – М.: Издательский дом "Вильямс", 2000.
5. *Ликеш И., Ляга И.* Основные таблицы математической статистики. – М.: Финансы и статистика, 1985.
6. *Мюллер П., Нойман П., Шторм Р.* Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
7. *Тьюки Дж.* Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981.

Зміст