РОЗДІЛ 1. ПРАВИЛО 3-СІГМА

Нерівність Чебишова-Б’єнаме стверджує, що для довільної випадкової величини x має місце співвідношення

$$p\left(\left|x-m(x)\right|\geq\lambda\right)\leq\frac{D(x)}{\lambda^2}.$$

Цю нерівність вперше сформулював (для часткового випадку) французький математик Ірене-Жюль Б’єнаме (Irénée-Jules Bienaymé, 1796–1878) в 1853 році, а довів (у загальному вигляді) — його друг і колега російський математик Пафнутій Львович Чебишо́в (1821-1894) у 1867 році.

У 1877 році німецький геодезист Вільгельм Йордан (1842–1899) наблизив стандартизований нормальний розподіл поліномом

$$\psi(x)=a+bx^2+cx^4,\ \ |x|<M$$

і, вважаючи, що $\psi'(M)=0$, обчислив всі три його параметри у функції. Під час обрахунків він помітив, що квадрат математичного сподівання пов’язаний з числом M наступним чином.

$$m^2=2\int_0^M x^2\psi(x)\,dx=\frac{M^2}{7}\approx\frac{M^2}{2{,}65^2}$$

Отже, $M\approx 2{,}65$. Виходячи з цього Йордан рекомендував відкидати всі спостереження, що відхиляються від свого середнього більше ніж на $3m$. Це і є знамените *правило 3-сігма.*

$$P\left(\left|x-m(x)\right|\geq 3\sigma(x)\right)\leq 0{,}05 \qquad (1.1)$$

Зауважимо, що французький математики Жан Батіст Жозеф Фур'є (1768-1830) у 1826 році сформулював аналогічне правило, щоправда, він пропонував збільшити цю величину в $\sqrt{2}$ рази. Це правило більше сторіччя залишалося емпіричним спостереженням, поки в 1980 році не було доведено нерівність Височанського-Петуніна, в основі якого лежить нерівність Гауса.

У 1821 році математик Карл Фрідріх Гаус (*Johann Carl Friedrich Gauß*; 1777–1855) довів наступну нерівність:

$$p\left(\left|x - M\left(x\right)\right| \geq k\tau\right) \leq \frac{4}{9k^2},$$

де $\tau^2 = Dx + \left(m(x) - M\left(x\right)\right)^2 > 0$ — другий момент розподілу $F(x)$ відносно моди M. Використаємо цю нерівність, щоб отримати строге доведення правила 3σ для випадкових величин із унімодальним розподілом.

Теорема (нерівність Височанського-Петуніна, 1980). *При всіх* $k > 0$ *для довільної випадкової величини x, що має унімодальний розподіл і скінченну дисперсію* $\sigma^2(x) > 0$ *має місце нерівність*

$$p\left(\left|x - m(x)\right| \geq k\sigma(x)\right) \leq \max\left\{\frac{4}{9} \cdot \frac{1}{k^2}, \frac{4-k^2}{3k^2}\right\} \tag{1.2}$$

Наслідок. *Для довільної випадкової величини x із унімодальним розподілом і скінченою дисперсією* $\sigma^2(x) > 0$ *при всіх* $k \geq \sqrt{\frac{8}{3}}$ *виконується така нерівність:*

$$p\left(\left|x - m(x)\right| \geq k\sigma(x)\right) \leq \frac{4}{9k^2} \tag{1.3}$$

Зокрема, для k=3

$$p\left(\left|x - m(x)\right| \geq 3\sigma(x)\right) \leq \frac{4}{81} \approx 0,049.$$

РОЗДІЛ 2. СТРУКТУРНА МОДЕЛЬ ВИПАДКОВОГО ЕКСПЕРИМЕНТУ

Розглянемо випробування T, що має два наслідки: A та $\overline{A}$. Введемо величину

$$x_k = \begin{cases} 1, \textit{якщо при } k - \textit{му повторенні випробування } T \textit{ відбувається подія } A, \\ 0 \ \textit{у протилежному випадку}. \end{cases}$$

Означення 2.3. Числова послідовність $x_1, x_2, \ldots$, що складається з нулів і одиниць, називається бернуліївською послідовністю порядку p, $0 \le p \le 1$, якщо

$$\lim_{n\to\infty} h_n(T, A) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k = p,$$

де $h_n(T, A)$ — частота подій A при n повтореннях випробування T.

Означення 2.4. Матриця

$$\Theta(T) = \begin{pmatrix} x_{11} & \ldots & x_{1n} & \ldots \\ x_{21} & \ldots & x_{2n} & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & \ldots & x_{nn} & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

називається характеризаційною матрицею випробування T.

Випадкове випробування, випадкова подія і ймовірність

Кожний рядок X_n і кожний стовпчик X_n^* матриці $\Theta(T)$породжують деякі дійсні числа α_n і α_n^* з відрізка $[0,1]$. Покладемо $\alpha_n = 0, x_{n1}x_{n2}...x_{nn}...$ і $\alpha_n^* = 0, x_{1n}x_{2n}...x_{nn}...$. Позначимо через M та M^* множини чисел α_n і α_n^*, відповідно.

Означення 2.5. Недетерміноване випробування T називається випадковим, якщо

1) всі рядки X_n і стовпчики X_n^* $(n=1,2,...)$ характеризаційної матриці $\Theta(T)$ є бернулієвськими послідовностями одного і того ж порядку $p \in [0,1]$;

2) множини чисел M та M^*, породжені рядками та стовпчиками характеризаційної матриці $\Theta(T)$, відповідно, є щільними на відрізку $[0,1]$.

Означення 2.6. Випадковим експериментом E будемо називати нескінченну серію повторень випадкового випробування T.

Означення 2.7. Випадковою подією (E,R) будемо називати результат R випадкового випробування T, що породжує випадковий експеримент E.

Означення 2.8. Ймовірністю $p(E,A)$ випадкової події (E,A) будемо називати порядок $p \in [0,1]$ бернулієвської послідовності, що складається із результатів випадкового випробування T, яке породжує випадковий експеримент E.

Теорема 2.1. Ймовірність того, що в схемі Бернуллі множини M і M^*, утворені рядками і стовпчиками характеризаційної матриці $\Theta(T)$ є скрізь щільними, дорівнює одиниці.

Поле подій

Нехай T — випадкове випробування, а $S(T)$— множина усіх випадкових подій, що можуть відбуватися внаслідок реалізації випробування T. Якщо у множині $S(T)$ для кожної події A визначено його ймовірність $p(A)$, то $S(T)$ ми називатимемо полем подій $S(E)$.

В полі подій $S(E)$ введемо відношення напівупорядкованості: $A \le B$, якщо поява події A в результаті випадкового експерименту E неодмінно тягне за собою появу події B. Це відношення перетворює поле подій $S(E)$ в напівупорядковану множину.

Для будь-яких двох подій $A, B \in S(E)$ мають місце співвідношення: $A + B = \sup(A, B) = A \vee B$ і $AB = \inf(A, B) = A \wedge B$,

Ці формули можна узагальнити для довільної множини подій:

$$\sum_{i \in J} A_i = \sup_{i \in J}\{A_i\} \text{ і } \prod_{i \in J} A_i = \inf_{i \in J}\{A_i\}.$$

Теорема 2.2. Для довільного випадкового експерименту E поле подій $S(E)$ є цілком дистрибутивною повною булевою алгеброю.

Означення 2.11. Називатимемо множину подій $B = \{B_i\}_{i \in J}$ із поля подій $S(E)$ базовою, якщо виконуються такі умови:

1) всі події B_i із B попарно несумісні $B_i B_j = 0$, якщо $i \neq j$;
2) довільну подію A з $S(E)$ можна подати у вигляді $A = \sum_{k \in K} B_{i_k}$.

Випадкові величини

Випадкову величину x визначимо, як функцію, визначену на базовій множині $B(E) \subset R^1$ випадкового експерименту E, при цьому у ролі атомів напівупорядкованого простору $S(E)$ виступають окремі значення випадкової величини x.

Теорема 2.4. Нехай $F(u)$ — довільна неперервна функція розподілу на прямій R^1, тоді існує випадковий експеримент E з базовою множиною $B_E = Q$ і розподілом ймовірностей $p(E, A)$, $A \subset Q$, для якого при будь-яких $u \in R^1$

$$p\left(E, Q_{(-\infty,u]}\right) = F(u), \quad \text{де } Q_{(-\infty,u]} = Q \cap (-\infty, u].$$

Теорема 2.5. Нехай $F(u)$ — довільна функція розподілу, зосереджена на відрізку $[a,b]$:

$$F(u) = \begin{cases} 0, \textit{ якщо } u \le a, \\ 1, \textit{ якщо } u \ge b. \end{cases}$$

Тоді існує випадковий експеримент E з числовою базовою множиною $B_E = [a,b]$, що породжує розподіл ймовірностей $p(E, A)$ на всіх підмножинах $A \subset [a,b]$, для якого за умови, що $u \in (a,b)$, має місце співвідношення

$$p\big(E, (a,u]\big) = F(u).$$

Операції над випадковими величинами

Інтерпретуючи випадкову величину як функцію, визначену на множині елементарних наслідків B_T деякого випадкового випробування T, можна ввести наступні операції:

$$(x+y)(B_c)=x(B_1)+y(B_2), \qquad (x-y)(B_c)=x(B_1)-y(B_2),$$

$$(xy)(B_c)=x(B_1)\,y(B_2), \qquad \left(\frac{x}{y}\right)(B_c)=\frac{x(B_1)}{y(B_2)},$$

де $x(B)$, $B\in B_{T_1}$ і $y(B)$, $B\in B_{T_2}$ — випадкові величини, що набувають свої значення в результаті проведення випадкових випробувань T_1 і T_2, $T_c=\{T_1,T_2\}$ — складене випробування, $B_c=(B_1,B_2)$.

Означення 2.19. Випадкові експерименти E_1 і E_2 називаються комутативними, якщо

$$p\big((E_1,E_2),(A_1,A_2)\big)=p\big((E_2,E_1),(A_2,A_1)\big).$$

Означення 2.20. Нехай E_1 і E_2 — два випадкових експерименти і $S(E_1)$, $S(E_2)$ — поле випадкових подій, породжене цими випадковими експериментами, відповідно. Будемо називати поля подій $S(E_1)$ і $S(E_2)$ ізоморфними, якщо між їх елементами можна установити взаємно однозначну відповідність Ψ, яка є структурним ізоморфізмом булевих алгебр $S(E_1)$ і $S(E_2)$ і зберігає ймовірність відповідних подій $p(E_1,A)=p\big(E_2,\Psi(A)\big)$, при цьому самі випадкові експерименти E_1 і E_2 також називаються ізоморфними.

Означення 2.21. Взаємно однозначна відповідність $\Psi: S(E_1) \to S(E_2)$, що є ізотонною відносно структур упорядкування, введених в булевих алгебрах $S(E_1)$ і $S(E_2)$, і володіє властивістю оберненої ізотонності, що зберігає ймовірність подій, тобто $\forall A \in S(E_1)$ $p(E_1, A) = p(E_2, \Psi(A))$, називається імовірнісним ізоморфізмом.

Теорема 2.6. Поля подій $S(E_c)$ і $S(\tilde{E}_c)$, породжені складеними експериментами $E_c = \{E_1, E_2\}$ і $\tilde{E}_c = \{E_2, E_1\}$ є ізоморфними, якщо випадкові експерименти E_1 і E_2 є комутативними.

Випадкова величина $x+y$ інтерпретується як результат випадкового експерименту E_c з базовою числовою множиною $B_{E_c} = B_{E_1} \oplus B_{E_2} = \left\{x+y : x \in B_{E_1}, x \in B_{E_2}\right\}$, де $B_{E_i}, i = 1,2$ — базова числова множина випадкового експерименту E_i. Випадкова величина $y+x$ ототожнюється із результатом випадкового експерименту $\tilde{E}_c$ із базовою множиною $B_{\tilde{E}_c} = B_{E_2} \oplus B_{E_1} = B_{E_1} \oplus B_{E_2} = B_{E_c}$. Поля подій $S(E_c)$ і $S(\tilde{E}_c)$ є ізоморфними, отже випадкові величини $x+y$ і $y+x$ можна ототожнити.

РОЗДІЛ 3. НЕПАРАМЕТРИЧНІ КРИТЕРІЇ ЕКВІВАЛЕНТНОСТІ НА ОСНОВІ МІР БЛИЗЬКОСТІ МІЖ ВИБІРКАМИ

Нехай $G=(M,F)$, $M\subset R^1$ – генеральна сукупність випадкових величин з невідомим розподілом імовірності F. **Гіпотеза Хілла**: якщо для одержання вибірки $(x_1,x_2,\ldots,x_n)\in G$ використовується випадковий вибір, такий що вибіркові значення $x_1,x_2,\ldots,x_n$ є симетрично залежними випадковими величинами з однаковим абсолютно неперервним розподілом, то

$$P\left(x_{n+1}=x\in\left(x_{(i)},x_{(j)}\right)\right)=\frac{j-i}{n+1},$$

де x_{n+1} – наступне вибіркове значення, а $x_{(i)}$, $x_{(j)}$— порядкові статистики.

Задача ідентифікації генеральної сукупності полягає в такому: нехай $x=(x_1,x_2,...,x_n)\in G$, $x'=(x'_1,x'_2,...,x'_m)\in G'$, треба з'ясувати: $z=(z_1,z_2,...,z_k)\in G$ або G'?

Для розв'язання цієї задачі необхідно

1) побудувати міру близькості $\rho_p(z,x)$ між вибірками z і x, для якої можна побудувати двосторонній довірчий інтервал $\left(\rho_p^{(l)},\rho_p^{(u)}\right)$, що відповідає заданому рівню значущості, не залежному від істинності чи хибності гіпотези H;
2) побудувати непараметричний критерій еквівалентності генеральних сукупностей, що дозволяє оцінити імовірність помилки 1-го і 2-го роду.

Міра близькості між вибірками

$H : F_G(x) \equiv F_{G'}(x)$, $x = (x_1, \dots, x_n) \in G$, $x' = (x'_1, x'_2, \dots, x'_m) \in G'$, $x_{(1)} \le \dots \le x_{(n)}$.

Позначимо $A_{ij}^{(k)} = \left\{ x_k \in \left(x_{(i)}, x_{(j)} \right) \right\}$. За гіпотезою Хілла, якщо $F_G(u) \equiv F_{G'}(u)$, то

$$P(A_{ij}^{(k)}) = P\left(x_k \in (x_{(i)}, x_{(j)}) \right) = p_{ij}^{(n)} = \frac{j-i}{n+1}.$$

За формулами Вільсона

$$p_{ij}^{(1)} = \frac{h_{ij}^{(n,m)} m + g^2/2 - g\sqrt{h_{ij}^{(n,m)}(1-h_{ij}^{(n,m)})m + g^2/4}}{m+g^2},$$

$$p_{ij}^{(2)} = \frac{h_{ij}^{(n,m)} m + g^2/2 + g\sqrt{h_{ij}^{(n,m)}(1-h_{ij}^{(n,m)})m + g^2/4}}{m+g^2}$$

де $h_{ij}^{(n,m)}$ – частота появи події $A_{ij}^{(k)}$ в m випробуваннях. Величина g визначає рівень значущості довірчого інтервалу $I_{ij} = \left(p_{ij}^{(1)}, p_{ij}^{(2)} \right)$; за правилом 3σ при $g = 3$ рівень значущості цього інтервалу не перевищує 0,05. Нехай $N = n(n-1)/2$, L – кількість інтервалів I_{ij}, які вміщують ймовірності p_{ij}, $h^{(N)} = \dfrac{L}{N}$ — міра близькості між вибірками x та x', або p-статистика.

Поклавши $h_{ij}^{(n,m)} = h^{(N)}$ та $g = 3$, отримуємо довірчий інтервал $I = \left(p^{(1)}, p^{(2)} \right)$ для ймовірності $p\left(p_{ij} \in I_{ij}^{(n,m)} \right)$. Довірчі інтервали $I_{ij} = \left(p_{ij}^{(1)}, p_{ij}^{(2)} \right)$ та $I = \left(p^{(1)}, p^{(2)} \right)$ будемо називати інтервалами, які побудовано за правилом 3σ.

Оцінка асимптотичного довірчого рівня значущості

Означення 3.1. Схема випробувань, при якій можуть з'являтися події $A_{ij}^{(k)}, k=1,2,...n$, у випадку істинності гіпотези *H*, називається узагальненою схемою Бернуллі.

Означення 3.2. Якщо гіпотеза *H* не є вірною, вказана схема випробувань називається модифікованою схемою Бернуллі.

Означення 3.3. У загальному випадку, коли може бути істинною будь-яка гіпотеза, як $F_G(u)=F_{G'}(u)$, так і $F_G(u)\neq F_{G'}(u)$, схема випробувань називається МП-схемою.

Теорема 3.2. Якщо в узагальненій схемі випробувань Бернуллі 1) $n=m$; 2) $0<\lim\limits_{n\to\infty} p_{ij}^{(n)} = p_0 <1$ та 3) $0<\lim\limits_{n\to\infty}\dfrac{i}{n+1} = p^* <1$, то асимптотичний рівень значущості β послідовності довірчих інтервалів $I_{ij}^{(n)}$ для ймовірностей $p_q^{(n)} = p_{ij}^{(n)} = p\left(A_{ij}^{(n)}\right)$, побудованих за правилом 3σ, не перевищує 0,05.

Означення 3.4. Нехай $B_1, B_2, ..., B_n, ...$ – послідовність подій, які можуть відбутися при проведенні випадкового експерименту *E* і $\lim\limits_{k\to\infty} p\left(B_k\right) = p^*$. Нехай $h_{n_1}\left(B_1\right), h_{n_2}\left(B_2\right), ..., h_{n_k}\left(B_k\right), ...$ – послідовність частот подій $B_1, B_2, ..., B_n, ...$, відповідно, причому $\dfrac{k}{n_k}\to 0$ при $k\to\infty$. Будемо називати експеримент *E* посиленим випадковим експериментом, якщо $h_{n_k}\left(B_k\right)\to p^*$ при $k\to\infty$.

Теорема 3.3. За умов посиленого випадкового експерименту E, якщо вибірки $x=\left(x_1,...,x_n\right)\in G$ та $x'=\left(x'_1,...,x'_m\right)\in G'$ мають однаковий об'єм, то асимптотичний рівень значущості інтервалу $I^{(n)}=\left(p^{(1)},p^{(2)}\right)$, побудований за правилом 3σ при $g=3$ за допомогою формул Вільсона, не перевищує 0,05.

Критерій для перевірки гіпотези про рівність функцій розподілу

Критерій 3.1. Якщо інтервал $I=\left(p^{(1)},p^{(2)}\right)$ містить 0,95, то приймається гіпотеза H, інакше приймається альтернативна гіпотеза.

Критерій для ідентифікації вибірки

Критерій 3.2. Нехай $x=\left(x_1,x_2,...,x_n\right)$ і $x'=\left(x'_1,x'_2,...,x'_n\right)$— вибірки, взяті з генеральних сукупностей G і G', відповідно, і $z=\left(z_1,z_2,...,z_n\right)$— вибірка, яка належить одній з цих двох генеральних сукупностей. Гіпотеза H_0 означає, що z взято з G, а $\overline{H}_0$– з G'. Позначимо через $h_{zx}\left(h_{zx'}\right)$ міру близькості між вибірками z і x (z і x'), через $I_{zx}=\left(p_{zx}^{(1)},p_{zx}^{(2)}\right)$ $\left(I_{zx}=\left(p_{zx}^{(1)},p_{zx}^{(2)}\right)\right)$ – відповідний довірчий інтервалі.

1) Якщо $p\left(p_{ij}^{(n)}\in I_{ij}^{(n,m)}\right)=0{,}95\in I_{zx}$ і $p_{zx'}^{(2)}<p_{zx}^{(1)}$, то приймається гіпотеза H_0;

2) Якщо $p\left(p_{ij}^{(n)}\notin I_{ij}^{(n,m)}\right)=0{,}95\in I_{zx'}$ і $p_{zx}^{(2)}<p_{zx'}^{(1)}$, то приймається гіпотеза $\overline{H}_0$;

3) У супротивному разі рішення про істинність гіпотез H_0 і $\overline{H}_0$ не приймається.

Ймовірності похибок 1-го і 2-го роду цього критерію не перевищують 0,05.

РОЗДІЛ 4. ДОВІРЧІ ІНТЕРВАЛИ З ГАРАНТОВАНИМ РІВНЕМ ЗНАЧУЩОСТІ

Розглянемо випадкову величину κ, яка дорівнює кількості елементів вибірки y, які потрапляють в інтервал $I_{i,q}$. Побудуємо точний довірчий інтервал $I(\kappa)$, що містить імовірність $P\left(\kappa = l \middle| H_0\right)$, використовуючи значення випадкової змінної κ.

Критерій 4.1.

1. Будуємо варіаційний ряд за вибіркою x і вибираємо випадковий інтервал $I_{i,q}$ для фіксованих i та q.
2. Визначаємо статистику Θ, що дорівнює кількості елементів вибірки y, які належать інтервалу $I_{i,q}$.
3. За допомогою статистики Θ будуємо довірчий інтервал $I(\Theta)$ із заданим рівнем значущості 2β.
4. Якщо інтервал $I_{i,q}$ містить $h = \dfrac{\theta}{m}$, ми робимо висновок, що дані не суперечать гіпотезі H_0. В протилежному випадку гіпотеза H_0 відхиляється на користь гіпотези H_1.

Обчислення показують, що при $n = m$ рівень значущості критерію 4.1 незначно відхиляється від 0,05.

Точний довірчий інтервал для невідомої ймовірності

Для побудови точних довірчих меж для невідомої ймовірності в моделях Бернуллі, розглянемо дві функції, що залежать від $p \in [0,1]$:

$$\varphi(p) = |h - p| \ \text{ і } \ \psi(p) = \frac{1}{2n} + \frac{\lambda}{n}\tilde{\psi}(p), \text{ де } \tilde{\psi}(p) = \sqrt{np(1-p) + \frac{1}{12}}, \ p \in R^1.$$

Нижня довірча межа p_1 є коренем квадратного рівняння

$$\left(1 + \frac{\lambda^2}{n}\right)p^2 - \left(\frac{\lambda^2}{n} - \frac{1}{n} + 2h\right)p + h^2 - \frac{h}{n} + \frac{1}{4n^2}\left(1 - \frac{\lambda^2}{3}\right) = 0. \qquad (4.1)$$

Якщо $h > \psi(0) = \frac{1}{2n} + \frac{\lambda}{n\sqrt{12}}$, то нижня довірча межа p_1 є найменшим коренем рівняння (4.1). Якщо $h \le \psi(0)$, то $p_1 = 0$.

Верхня довірча межа p_2 є коренем квадратного рівняння

$$\left(1 + \frac{\lambda^2}{n}\right)p^2 - \left(\frac{\lambda^2}{n} + \frac{1}{n} + 2h\right)p + h^2 + \frac{h}{n} + \frac{1}{4n^2}\left(1 - \frac{\lambda^2}{3}\right) = 0. \qquad (4.2)$$

Якщо $1 - h > \psi(1)$, то верхня довірча межа p_2 є найбільшим коренем рівняння (4.2). Якщо $1 - h \le \psi(1)$, то $p_2 = 1$.

Для узагальненої схеми Бернуллі отримаємо квадратне рівняння (4.3).

$$\left(1 + \frac{(m+n+1)\lambda^2}{(n+2)m}\right)p^2 + \left(\frac{1}{m} - \frac{(m+n+1)\lambda^2}{(n+2)m} - 2h\right)p + h^2 - \frac{h}{m} + \frac{1}{4m^2}\left(1 - \frac{\lambda^2}{3}\right) = 0. \qquad (4.3)$$

Якщо $h > \frac{1}{2m} + \frac{\lambda}{m\sqrt{12}} = \gamma$, то нижня довірча межа p_1 в узагальненій схемі Бернуллі є найменшим коренем рівняння (4.3). Якщо $h \le \gamma$, то $p_1 = 0$.

Верхня довірча межа p_2 в узагальненій схемі Бернуллі є коренем рівняння (4.4).

$$\left(1+\frac{(m+n+1)\lambda^2}{(n+2)m}\right)p^2 - \left(\frac{1}{m}+\frac{(m+n+1)\lambda^2}{(n+2)m}+2h\right)p + h^2 + \frac{h}{m} + \frac{1}{4m^2}\left(1-\frac{\lambda^2}{3}\right) = 0. \quad (4.4)$$

Якщо $1-h > \gamma$, то верхня довірча межа p_2 є найбільшим коренем рівняння (4.4). Якщо $1-h \le \gamma$, то $p_2 = 1$. Рівень значущості побудованих довірчих інтервалів не перевищує $\frac{4}{9}\frac{1}{\lambda^2}$.

Критерій 4.2

1. Будуємо на основі вибірки x варіаційний ряд і випадковий інтервал $I_{i,q} = \left(x_{(i)}, x_{(i+q)}\right)$, де i і q — фіксовані числа, $0 \le i \le n$, $1 \le q \le n-i+1$.
2. Обчислюємо статистику κ, що дорівнює кількості елементів із вибірки y, які потрапляють в інтервал $I_{i,q}$.
3. Будуємо точний довірчий інтервал $I(\kappa)$, рівень значущості якого дорівнює 2β.
4. Якщо інтервал $p_q = \frac{q}{n+1} \notin I(\kappa)$, гіпотеза H_0 відхиляється, інакше вона приймається.

Теорема 4.1. Рівень значущості критерію 4.2 в рамках класичної і узагальненої моделей Бернуллі дорівнює $\frac{4}{9}\frac{1}{\lambda^2}$, де $\lambda > \sqrt{\frac{8}{3}}$.

Порівняння невідомих ймовірностей

Друга задача формулюється так: побудувати статистичний критерій, що дозволяє по наявних частотах h_1 і h_2 зробити висновок про значуще відхилення ймовірностей p_1 і p_2, не використовуючи умову незалежності частот.

Нехай B — схема Бернуллі, в якій випадкова подія A може відбутися з імовірністю p, а частота h випадкової події A в n випробуваннях. Як відомо, $m(h)=p$, $D(h)=\dfrac{pq}{n}$, де $q=1-p$, зміщена оцінка $D(h)$: $s_n^2=\dfrac{h(1-h)}{n}$, незміщена оцінка $D(h)$: $\hat{s}_n^2=\dfrac{n}{n-1}s_n^2=\dfrac{h(1-h)}{n-1}$.

Теорема 4.2. При будь-якому натуральному $n\geq 2$

$$\Delta\left(s_n^2\right)=m\left(s_n^2-D(h)\right)^2<D\left(\hat{s}_n^2\right).$$

Нехай $I_1=\left(p_1^{(1)},p_2^{(1)}\right)$ і $I_2=\left(p_1^{(2)},p_2^{(2)}\right)$ — довірчі інтервали для невідомих ймовірностей p_1 і p_2, відповідно, рівень значущості яких приблизно дорівнює 5%.

Критерій 4.3. Основна гіпотеза H про рівність ймовірностей p_1 і p_2 ($p_1=p_2=p$) відхиляється, якщо довірчі інтервали I_1 і I_2 не перетинаються.

Теорема 4.3. Якщо основна гіпотеза H є вірною, то рівень значущості критерію 4.3 не перевищує 0,05.

Критерій 4.4. Гіпотеза H відхиляється, якщо середина інтервалу I_2 не належить інтервалу I_1.

Теорема 4.4. Якщо основна гіпотеза H є вірною, то рівень значущості критерію 4.4 не перевищує 0,05.

РОЗДІЛ 5 СТРАТИФІКАЦІЙНИЙ АНАЛІЗ ГЕНЕРАЛЬНИХ СУКУПНОСТЕЙ

Нехай G – генеральна сукупність з неперервною строго зростаючою функцією розподілу $F(u)$. Носієм функції розподілу $F(u)$ називається мінімальний відрізок $[a,b]$, для якого $F(u)\equiv 0\ \forall u\le a$ і $F(u)\equiv 1\ \forall u\ge b$. Розглянемо варіаційний ряд $x_{(1)}\le x_{(2)}\le\ldots\le x_{(n)}$, побудований по вибірці $x_1,x_2,\ldots,x_n$ і визначимо кусково-постійну емпіричну функцію розподілу $F_n^*(u)$. За теоремою Глівенка-Кантеллі $p\left(\lim\limits_{n\to\infty}\max\limits_u\left|F(u)-F_n^*(u)\right|=0\right)=1$.

Задача 5.1. Отримати таку ж оцінку для оберненої функції розподілу.

Лінійний сплайн. Для неперервної монотонно зростаючої функції розподілу $F(u)$ з компактним носієм $[a,b]$ побудуємо такий лінійний сплайн.

$$\tilde{F}_n^*(u)=\begin{cases}\dfrac{u}{n}+\dfrac{kx_{(k+1)}-(k+1)x_{(k)}}{n}, \textit{ якщо } x_{(k)}\le u\le x_{(k+1)},\ x_{(0)}=a,\ k=0,1,\ldots,n;\\ 1,\ \textit{якщо } x_{(n)}\le u\le b.\end{cases}$$

Кусково-лінійна емпірична функція розподілу $\tilde{F}_n^*(u)$, визначена на відрізку $[a,b]$, має на відрізку $\left[a,x_{(n)}\right]$ обернену функцію $\left(\tilde{F}_n^*(u)\right)^{-1}\triangleq\Psi_n^*(u),\ 0\le u\le 1$. Функція $\Psi_n^*(u)$ є неперервною кусково-лінійною функцією, що проходить через точки $\left(\dfrac{k}{n},x_{(k)}\right)$, де $x_{(k)}$ — порядкові статистики.

Лема 5.1. Якщо гіпотетична функція розподілу $F(u)$ генеральної сукупності є неперервною і має компактний носій $[a,b]$, то послідовність кусково-лінійних емпіричних функцій розподілу із ймовірністю одиниця збігається до функції розподілу в чебишовській метриці: $p\left(\lim_{n\to\infty}\max_{u}\left|F(u)-\tilde{F}_n^*(u)\right|=0\right)=1$.

Теорема 5.1. Якщо гіпотетична функція розподілу $F(u)$ генеральної сукупності G є неперервною і має компактний носій $[a,b]$, то послідовність функцій $\left\{\Psi_n^*(u)\right\}$ збігається із ймовірністю одиниця до функції $F^{-1}(u)$, тобто $p\left\{\lim_{n\to\infty}\max_{a\le u\le b}\left|F^{-1}(u)-\Psi_n^*(u)\right|=0\right\}=1$.

Задача 5.2. Побудувати критерій одномодальності функції розподілу $F(u)$ на підставі вибірки $x_1,x_2,...,x_n$, отриманої зі генеральної сукупності значень випадкової величини x за допомогою простого випадкового вибору.

Визначимо оцінку $h_n(u)$ для гіпотетичної щільності ймовірностей:

$$h_n(u)=\begin{cases}\dfrac{1}{(n-1)\left(x_{(i+1)}-x_{(i)}\right)}, \text{ якщо } u\in\left[x_{(i)},x_{(i+1)}\right),\\ 0, \text{ якщо } u<x_{(1)} \text{ або } u>x_{(n)}.\end{cases} \tag{5.1}$$

Функцію $h_n(u)$ називатимемо уніфікованою гістограмою.

Інтеграл від уніфікованої гістограми

$$\tilde{F}_n^*(u) = \int\limits_{x_{(1)}}^{u} h_n(v)\,dv = \frac{u + (i-1)x_{(i+1)} - ix_{(i)}}{(n-1)\left(x_{(i+1)} - x_{(i)}\right)}, \text{ якщо } x_{(i)} \le v < x_{(i+1)}, \tag{5.2}$$

являє собою лінійний сплайн.

Будемо називати функцію $\tilde{F}_n^*(u)$, визначену формулою (5.2), модифікованою емпіричною функцією розподілу.

Для заданого рівня значущості β^* за таблицею точних і асимптотичних меж для верхньої грані модуля різниці між справжньою і емпіричною функціями розподілу можна визначити величину ε, для якої

$$p\left(\Delta = \max_{x_{(1)} \le u \le x_{(n)}} \left|F(u) - F_n(u)\right| > \varepsilon\right) = \beta^*,$$

і побудувати смугу Π_{β^*}, межами якої служать східчасті лінії з рівняннями $y = F_n^*(u) + \varepsilon$ і $y = F_n^*(u) - \varepsilon$. Смуга Π_{β^*} цілком накриває справжню криву розподілу з рівнянням $y = F(u)$ з імовірністю (довірчим рівнем) $\alpha^* = 1 - \beta^*$. Називатимемо смугу Π_{β^*} довірчою смугою для емпіричної функції розподілу з рівнем значущості β^*, побудованою на підставі емпіричної функції розподілу.

Критерій одномодальності, заснований на емпіричній функції розподілу

Означення 5.1 Нехай $y = \varphi(u)$ — довільна функція, задана на відрізку $[a,b]$. Будемо називати надграфіком функції $\varphi(u)$ множину точок $G_U = \{(u, y) : y \geq \varphi(u), a \leq u \leq b\}$, а підграфіком — множину точок $G_L = \{(u, y) : y \leq \varphi(u), a \leq u \leq b\}$.

Означення 5.2. Будемо називати опуклою мінорантою $\varphi_{conv}(u)$ функції $\varphi(u)$ нижню межу опуклої оболонки її надграфіку $\varphi_{\inf}(u) = \inf\left\{v : (u,v) \in \underset{a \leq u \leq b}{conv} G_U\right\}$, де через $conv\, G_U$ позначена опукла оболонка множини G_U. Аналогічно будемо називати угнутою мажорантою функції $\varphi(u)$ верхню межу опуклої оболонки її підграфіку $\varphi_{\sup}(u) = \sup\left\{v : (u,v) \in \underset{a \leq u \leq b}{conv} G_U\right\}$.

Теорема 5.3 (критерій 5.1). Нехай $\varphi_{\inf}(u)$, де $x_{(1)} \leq u \leq x_{(n)}$ — опукла міноранта функції $\varphi(u)$ (верхньої межа смуги Π_{β^*}), а $\psi_{\sup}(u)$, де $x_{(1)} \leq u \leq x_{(n)}$ — угнута мажоранта функції $\psi(u)$ (нижньої межі смуги Π_{β^*}), і

$$\alpha = \sup\left\{u : \varphi_{\inf}(u) \leq \psi(u),\ x_{(1)} \leq u \leq x_{(n)}\right\},$$

$$\beta = \inf\left\{u : \psi_{\sup}(u) \geq \varphi(u),\ x_{(1)} \leq u \leq x_{(n)}\right\};$$

Гіпотетичний розподіл $F(u)$ є одномодальним, якщо і тільки якщо:

1) $\varphi_{\inf}(u) \geq \psi(u)$ або $\varphi_{\sup}(u) \leq \varphi(u)$ $\forall u \in \left[x_{(1)}, x_{(n)}\right]$; або 2) $\alpha \geq \beta$.

При цьому рівень значущості цього критерію дорівнює β^*.

Критерій одномодальності, заснований на модифікованої емпіричній функції розподілу

Побудуємо довірчу смугу $\tilde{\Pi}_{\beta^*}$ для гіпотетичної функції розподілу за допомогою модифікованої емпіричної функції розподілу $\tilde{F}_n^*(u)$. Нехай β^* — заданий рівень значущості, а $\varepsilon > 0$ — ширина довірчої смуги Π_β, тоді

$$p\left(\Delta = \max_{x_{(1)} \le u \le x_{(n)}} \left|F(u) - F_n^*(u)\right| > \varepsilon\right) = \beta^*.$$

Покладемо $\tilde{\varphi}(u) = \tilde{F}_n^*(u) + \varepsilon + \frac{1}{n}$, $\tilde{\psi}(u) = \tilde{F}_n^*(u) - \varepsilon - \frac{1}{n}$.

Теорема 5.4 (критерій 5.2) Нехай $\tilde{\varphi}_{\inf}(u)$, де $x_{(1)} \le u \le x_{(n)}$ — опукла мінoранта функції $\tilde{\varphi}(u)$ (верхньої межі смуги $\tilde{\Pi}_{\beta^*}$), а $\tilde{\psi}_{\sup}(u)$, де $x_{(1)} \le u \le x_{(n)}$ — угнута мінoранта функції $\tilde{\psi}(u)$ (нижньої межі смуги $\tilde{\Pi}_{\beta^*}$), і

$$\alpha = \sup\left\{u : \tilde{\varphi}_{\inf}(u) \le \tilde{\psi}(u),\ x_{(1)} \le u \le x_{(n)}\right\},$$
$$\beta = \inf\left\{u : \tilde{\psi}_{\sup}(u) \ge \tilde{\varphi}(u),\ x_{(1)} \le u \le x_{(n)}\right\};$$

Гіпотетичний розподіл $F(u)$ є одномодальним, якщо і тільки якщо виконується одна з двох умов:

1) $\tilde{\varphi}_{\inf}(u) \ge \tilde{\psi}(u)$ або $\tilde{\varphi}_{\sup}(u) \le \tilde{\varphi}(u)$ $\forall u \in \left[x_{(1)}, x_{(n)}\right]$;

2) $\alpha \ge \beta$.

При цьому рівень значущості цього критерію дорівнює β^*.