ЛЕКЦІЯ 4
ГРУБІ ПОМИЛКИ. ВІДКИДАННЯ ДАНИХ. ОПТИМІЗАЦІЯ ЗАСМІЧЕННЯ ВИБІРКИ

Груба помилка (промах, викид) – це похибка результату окремого вимірювання, що входить до ряду вимірювань, яка для даних умов дуже відрізняється від результатів цього ряду.

Іноді результат одного з серії вимірювань разюче розходиться з іншими. Коли це відбувається, експериментатор повинен вирішити, чи є такий аномальний результат вимірювання наслідком деякої грубої помилки і тому повинен бути відкинутий, або ж це коректний результат, який повинен розглядатися разом з іншими.

Відкидання даних – важливе та спірне питання, щодо якого у фахівців немає єдиної думки. Рішення відкинути якісь дані є, зрештою, завжди суб'єктивним.

Джерелом грубих помилок можуть бути різкі зміни умов вимірювання і помилки, допущені дослідником. До них можна віднести вихід з ладу приладу або поштовх, неправильний відлік за шкалою вимірювального приладу, неправильний запис результату досліджень, хаотичні зміни параметрів напруги, що живить засіб вимірювання тощо.

В одних випадках грубі помилки одразу помітні серед отриманих результатів, оскільки вони сильно відрізняються від інших значень. Проте треба відзначити складність виявлення грубих помилок, пов’язану з так званим "***маскуючим ефектом***". Результати, які підозрюються в аномальності, часто групуються близько один до одного, створюючи групу, яка дещо відстроїть від основної маси результатів. Це робить послідовні процедури нечутливими до них.

Наявність грубих помилок може сильно спотворити результат експерименту. Але необдумане відкидання результатів, які різко відрізняються від інших, може також призвести до суттєвого спотворення

характеристик вимірювань. Тому первинна обробка експериментальних даних рекомендує будь-яку сукупність вимірювань перевіряти на наявність грубих помилок за допомогою ***статистичних критеріїв***.

Відомо, що при проведенні серії вимірювань результати окремих вимірювань x_i розміститься поблизу невідомого істинного значення x таким чином, що їх відхилення в бік більших або менших значень будуть рівноімовірними. При цьому найкращим наближенням до істинного значення є середнє значення $\bar{x}$ з N вимірювань.

Результат вимірювання прийнято вказувати у вигляді ***довірчого інтервалу*** значень вимірюваної величини, в межах якого з визначеною імовірністю визначається істинне значення x. Для довірчого інтервалу обов'язково вказують кількісну характеристику його достовірності – ***довірчу імовірність*** P. Під ***довірчою імовірністю*** зазвичай розуміють відношення кількості дослідів, що дають вказаний в інтервалі результат, до загального числа проведених дослідів, *або* імовірність того, що істинне значення вимірюваної величини визначається всередині довірчого інтервалу, поблизу отриманого середнього значення.

Довірчий інтервал повинен перекривати максимальні імовірності розподілу, а сама оцінка – бути всередині. ***Рівень значущості*** (він дорівнює $1-P$) – це імовірність того, що сумнівний результат опиниться за межами довірчого інтервалу. Так, при довірчій імовірності 0.95, рівень значущості дорівнює 0.05 тощо.

З лекції 3 пам'ятаємо розповсюджений спосіб запису результату вимірювань за допомогою довірчого інтервалу: $x = \bar{x} \pm \sigma_{\bar{x}}$,

де $\sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{N}}$ – стандартне відхилення середнього;

$\sigma_x = \sqrt{\dfrac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ – стандартне відхилення.

Якщо результати наших вимірювань розподілені нормально, і якщо ми повторимо вимірювання x дуже багато разів (завжди з однаковою

апаратурою), то приблизно 70% (*точніше 68.28%*) результатів наших вимірювань перебуватимуть в межах $\sigma_{\bar{x}}$ від $\bar{x}$ (70% результатів наших вимірювань належатимуть інтервалу $\bar{x} \pm \sigma_{\bar{x}}$). Тобто довірчій імовірності P=0.68 відповідає довірчий інтервал, визначений формулою $\sigma_{\bar{x}} = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N(N-1)}}$. Очевидно, що збільшення числа вимірювань N призводить до зменшення похибки результату $\sigma_{\bar{x}}$.

Повернемося до питання, чи відкидати результат вимірювань, який здається до такої міри некоректним, що швидше схожий на помилку.

Зазвичай перевіряють найбільше та найменше значення результатів вимірювань.

Наприклад, нехай ми робимо шість вимірювань періоду коливань маятника і отримуємо такі результати (в секундах):

3.8; 3.5; 3.7; 3.9; 3.4; 1.8. (4.1)

Тут значення 1.8 значно відрізняється від інших, і ми повинні вирішити, що з ним робити.

В нашому прикладі найкраща оцінка періоду коливань маятника принципово залежить від того, чи відкинемо ми підозріле значення 1.8 с. Середнє всіх шести вимірювань дорівнює 3.4 с, в той час як середнє п’яти вимірювань (без 1.8) дорівнює 3.7 с, тобто суттєво відрізняється.

З іншого боку, відкидаючи значення часу 1.8 с, ми, можливо, відкидаємо найцікавішу частину даних. Першою реакцією на подібні до (4.1) дані є багаторазове повторення вимірювань. Якщо аномалія повториться, ми, імовірно, зможемо визначити її причину, є то помилкою або реальним фізичним ефектом; якщо ж вона не повториться, то до того часу ми зробимо, скажімо, 100 вимірювань, так що не буде суттєвої різниці для нашого кінцевого результату, врахуємо ми аномалію чи ні. Однак в більшості

випадків непрактично повторювати вимірювання 100 разів, якщо результат здається підозрілим.

Отже, необхідний критерій, згідно з яким виключається (відкидається) підозрілий результат.

Існує багато різних критеріїв, кожен з яких застосовується у відповідних випадках. Іноді корисно проводити оцінювання за кількома критеріями.

4.1 Виключення грубих помилок

4.1.1 Критерій Діксона

Варіаційний ***критерій Діксона*** *ґрунтується на припущенні, що результати вимірювань підпадають під нормальний закон розподілу. Він вважається надійним* ***при малому числі вимірювань N≤10***. При його застосуванні результати вимірювань впорядковують і записують у вигляді варіаційного зростаючого ряду: $x_1<x_2<\ldots<x_N$. В цьому випадку *результат вважається грубою помилкою*, якщо

$$K_d = \frac{x_N - x_{N-1}}{x_N - x_1} > Z_d\,, \qquad (4.2)$$

де K_d – розраховане значення критерію Діксона;

Z_d – критичне значення критерію Діксона залежно від розміру вибірки N і довірчої ймовірності P (див. табл.4.1).

Застосуємо критерій Діксона до вибірки (4.1). Розміщуємо результати досліджень (4.1) у вигляді варіаційного зростаючого ряду: 1.8; 3.4; 3.5; 3.7; 3.8; 3.9. Підставляємо у (4.2) дані (4.1) та розраховуємо K_d: $K_d = \frac{3.9-3.8}{3.9-1.8} = 0.048$. Як видно з табл.4.1, для всіх довірчих імовірностей $K_d < Z_d$. Отже, за критерієм Діксона, ряд результатів досліджень (4.1) не має у своєму складі грубих помилок навіть при 10%-вому рівні значущості. Тому під подальшу обробку підпадатиме весь масив даних.

Таблиця 4.1 – Критичні значення критерію Діксона

N	Z_d			
	0.10 (*P*=0.90)	0.05 (*P*=0.95)	0.02 (*P*=0.98)	0.01 (*P*=0.99)
4	0.68	0.76	0.85	0.89
5	0.56	0.64	0.73	0.78
6	0.48	0.56	0.64	0.70
7	0.43	0.51	0.60	0.64
8	0.40	0.47	0.54	0.59
9	0.37	0.44	0.51	0.56
10	0.35	0.41	0.48	0.53
12	0.32	0.38	0.44	0.48
14	0.29	0.35	0.41	0.45
16	0.28	0.33	0.39	0.43
18	0.26	0.31	0.37	0.41
20	0.26	0.30	0.36	0.39
25	0.23	0.28	0.33	0.36
30	0.22	0.26	0.32	0.34

4.1.2 Критерій Романовського

Критерій Романовського *надійний* ***при*** $N \le 20$. Значення x_i вважають грубою помилкою, якщо виконується нерівність

$$\frac{|\bar{x} - x_i|}{\sigma_x} \ge \beta_r, \qquad (4.3)$$

де $\bar{x} = \frac{\sum x_i}{N}$ – середнє значення (математичне очікування);

$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ – стандартне відхилення;

β_r – критичне значення критерія Романовського залежно від розміру вибірки *N* та довірчої імовірності *P* (див. табл.4.2).

При цьому величини $\bar{x}$ *та* σ_x *обчислюються без урахування екстремальних (підозрілих) значень*.

Для критерія Романовського зазвичай обирають рівень значущості 0.01–0.05, тобто високу довірчу імовірність P.

Таблиця 4.2 – Критичні значення критерія Романовського

N	β_r			
	0.10 (P=0.90)	0.05 (P=0.95)	0.02 (P=0.98)	0.01 (P=0.99)
4	1.69	1.71	1.72	1.73
6	2.00	2.10	2.13	2.16
8	2.17	2.27	2.37	2.43
10	2.29	2.41	2.54	2.62
12	2.39	2.52	2.66	2.75
15	2.49	2.64	2.80	2.90
20	2.62	2.78	2.96	3.08

Застосуємо критерій Романовського до вибірки (4.1), щоб визначити, чи є грубою помилкою значення 1.8 с. Без урахування сумнівного значення вибірка (4.1) має вигляд: 3.8; 3.5; 3.7; 3.9; 3.4. Обчислюємо $\bar{x}$ та σ_x для нової вибірки: $\bar{x} = \dfrac{3.8+3.5+3.7+3.9+3.4}{5} = 3.7\,(\text{с}),$

$$\sigma_x = \sqrt{\frac{(3.8-3.7)^2+(3.5-3.7)^2+(3.7-3.7)^2+(3.9-3.7)^2+(3.4-3.7)^2}{5-1}} = 0.21(\text{с}),$$

$\dfrac{|3.7-1.8|}{0.21} = 9.05 > \beta_r$ для всіх значень довірчої імовірності (див. табл.4.2). Отже, за критерієм Романовського, значення 1.8 є грубою помилкою і його потрібно відкинути.

4.1.3 Критерій 3σ

Критерій 3σ *застосовується для результатів вимірювань, розподілених за нормальним законом. Даний критерій надійний* ***при числі вимірювань $20 \le N \le 50$****. Середнє значення $\bar{x}$ і стандартне відхилення σ_x обчислюються <u>без урахування</u> екстремальних (підозрілих) значень x_i.* У цьому випадку *<u>грубою помилкою вважається результат x_i</u>*, якщо

$$|\bar{x} - x_i| > 3\sigma_x. \qquad (4.4)$$

Це означає, що імовірність того, що випадкова величина відхилиться від свого математичного очікування на більшу величину, ніж потроєне середнє квадратичне відхилення, практично дорівнює нулю.

Перевірку результатів проводять послідовно, виключаючи з вибірки грубі помилки доти, поки не виконуватиметься нерівність (4.4). *Правило* 3σ *вважається занадто жорстким, тому рекомендується, залежно від розміру вибірки N, використовувати більш м'які критерії:*

- *при* $6 < N \le 100$ *– правило* $4\sigma_x$;
- *при* $100 < N \le 1000$ *– правило* $4.5\sigma_x$; (4.5)
- *при* $1000 < N \le 10000$ *– правило* $5\sigma_x$.

4.1.4 Критерій Шовене

Критерій Шовене – *простий випадок застосування розподілу Гауса. Його застосовують,* ***якщо число вимірювань*** $N \le 10$. У цьому випадку грубою помилкою вважається результат x_i, якщо залежно від числа вимірювань (розміру вибірки) N виконується нерівність:

$$|\bar{x} - x_i| > \begin{cases} 1.6\sigma_x \text{ при } N = 3; \\ 1.7\sigma_x \text{ при } N = 6; \\ 1.9\sigma_x \text{ при } N = 8; \\ 2\sigma_x \text{ при } N = 10. \end{cases} \quad (4.6)$$

Сумнівне значення приймає участь у розрахунку характеристик x_i *і* σ_x.

Повернемося до нашого прикладу. З урахуванням сумнівного значення 1.8 с для вибірки (4.1) обчислюємо $\bar{x}$ та σ_x: $\bar{x}$ =3.4 с,

$$\sigma_x = \sqrt{\frac{(3.8-3.4)^2 + (3.5-3.4)^2 + (3.7-3.4)^2 + (3.9-3.4)^2 + (3.4-3.4)^2 + (1.8-3.4)^2}{6-1}} =$$

=0.78 (с). Тоді за (4.6) для N=6 $|\bar{x} - x_i| = 1.6 > 1.7 \cdot 0.78 = 1.3$. Отже, за критерієм Шовене, 1.8 с – груба помилка, яку потрібно відкинути.

Після відкидання результату, що не задовольняє критерій Шовене, слід перерахувати $\bar{x}$ і σ_x за залишеними даними, щоб оцінити, наскільки в

реальності підозрюване значення впливає на кінцевий результат. Таким чином, для вибірки 3.8; 3.5; 3.7; 3.9; 3.4 $\bar{x}$ =3.7 с і σ_x=0.21 с. Як видно з розрахунків, середнє значення змінилося ненабагато, а стандартне відхилення суттєво зменшилось.

4.1.5 Критерій Грабса

Критерій Грабса *застосовується* ***для вимірювань середнього обсягу N<50.*** В цьому випадку грубою помилкою вважається результат x_i, якщо виконується така нерівність

$$\frac{|\bar{x} - x_i|}{\sigma_x} \geq G, \qquad (4.7)$$

де *G* – табличне (критичне) значення відсоткових точок критерія Грабса (див. табл.4.3) для відповідної довірчої ймовірності *P*. Якщо сумнівне значення є помилкою, воно повинне бути виключене з даних експерименту.

Недоліками критерія Грабса є його неточність і нечутливість до засмічень, коли помилки об'єднуються у групи на відстані від загальної сукупності.

Для підозрілого значення 1.8 с з вибірки (4.1) 3.8; 3.5; 3.7; 3.9; 3.4; 1.8 $\bar{x}$ =3.4 с, σ_x=0.78 с. Тоді для *N*=6: $\frac{|\bar{x} - x_i|}{\sigma_x} = \frac{|3.4 - 1.8|}{0.78} = 2.051$. Це значення більше табличних значень (табл.4.3) для довірчих імовірностей 0.90 та 0.95 і в цих випадках значення 1.8 с з вибірки слід виключити. У випадках же двох- та 1%-вого рівнів значущості отримане значення менше, ніж відповідні табличні, і, отже, за критерієм Грабса, значення 1.8 с з вибірки не слід виключати.

Таблиця 4.3 – Відсоткові точки критерія Грабса

N	*G*			
	0.10 (*P*=0.90)	0.05 (*P*=0.95)	0.02 (*P*=0.98)	0.01 (*P*=0.99)
3	1.406	1.412	1.414	1.414
4	1.645	1.689	1.710	1.723
5	1.791	1.869	1.917	1.955
6	1.894	1.996	2.067	2.130
7	1.947	2.093	2.182	2.265
8	2.041	2.172	2.273	2.374
9	2.097	2.238	2.349	2.464
10	2.146	2.294	2.414	2.540
11	2.190	2.343	2.470	2.606
12	2.229	2.387	2.519	2.663
13	2.264	2.426	2.563	2.713
14	2.297	2.461	2.602	2.759
16	2.354	2.523	2.670	2.837
18	2.404	2.577	2.728	2.903
20	2.447	2.623	2.779	2.959
22	2.486	2.664	2.823	3.008
24	2.521	2.701	2.862	3.051
26	2.553	2.734	2.897	3.089
28	2.582	2.764	2.929	3.124
30	2.609	2.792	2.958	3.156
35	2.668	2.853	3.022	3.224
40	2.718	2.904	3.075	3.281
45	2.762	2.948	3.120	3.329
50	2.800	2.987	3.160	3.370

4.1.6 Критерій Ірвіна

Якщо розподіл результатів досліджень не є нормальним або невідомий, то для оцінки підозрілих результатів можна використати ***критерій Ірвіна***, згідно з яким підозріле значення вважається аномальним, якщо воно відрізняється від попереднього на величину, більшу за середньоквадратичне відхилення σ_x. При цьому будують варіаційний ряд значень (за збільшенням або зменшенням) і оцінюють підозрілі значення на одному або обох краях ряду. Для цього обчислюють розрахункове значення критерію Ірвіна:

$$\eta_i = \frac{|x_i - x_{i-1}|}{\sigma_x}, \quad i = \overline{2, N}. \tag{4.8}$$

При розрахунку характеристик сумнівне значення враховується.

Розрахункові значення $\eta_2, \eta_3, ..., \eta_N$ порівнюються з табличними значеннями критерію Ірвіна η_α (див. табл.4.4) і, *якщо вони виявляються більшими за табличні, то відповідне значення* x_i *вважають аномальним, відкидають і перевіряють наступне*. Перевірку продовжують доти, поки не отримують $\eta_i < \eta_\alpha$.

Таблиця 4.4 – Критичні значення критерію Ірвіна

Довірча імовірність P	η_α
0.90	$2N^{-0.5}+0.6$
0.95	$2.5N^{-0.5}+0.75$
0.99	$3N^{-0.5}+1.15$

Табличне значення критерію Ірвіна розраховують з прийнятною точністю за залежностями, наведеними в табл.4.4 (при кількості випробувань (дослідів) N в межах від 3 до 1000).

Для вибірки (4.1) розглянемо необхідність відкидання значення 1.8 с за критерієм Ірвіна. Для зручності розмістимо значення ряду за зменшенням: 3.9; 3.8; 3.7; 3.5; 3.4; 1.8. Визначаємо значення $\bar{x}$ =3.4, σ_x =0.78 с, розрахункове значення критерію Ірвіна $\eta_6 = \dfrac{|x_6 - x_5|}{\sigma_x} = \dfrac{|1.8 - 3.4|}{0.78} = 2.1$, а також критичні значення критерію Ірвіна для різних довірчих імовірностей:

$$\eta_\alpha^{P=0.90} = 2 \cdot 6^{-0.5} + 0.6 = 1.4 < \eta_6 = 2.1,$$

$$\eta_\alpha^{P=0.95} = 2.5 \cdot 6^{-0.5} + 0.75 = 1.8 < \eta_6 = 2.1,$$

$$\eta_\alpha^{P=0.99} = 3 \cdot 6^{-0.5} + 1.15 = 2.4 > \eta_6 = 2.1.$$

Як показують розрахунки, за критерієм Ірвіна, для довірчих імовірностей 0.90 і 0.95 значення вибірки 1.8 с є аномальним і повинне бути відкинуте. У випадку ж 1%-вого рівня значущості (P=0.99) підозріле значення 1.8 не є аномальним і не відкидається.

Наведені вище методи попередньої обробки результатів вимірювань зазвичай ефективні при нормальному законі розподілу випадкових похибок виправленого (без значної систематичної складової) результату.

4.2 Робастні методи

Часто у дослідника відсутні достатні підстави вважати закон розподілу нормальним, і, до того ж, досить часто на практиці реалізуються інші закони розподілу. В цих випадках середнє значення вже не є оптимальною оцінкою вимірюваного значення величини, а дисперсія емпіричного розподілу, отримана відомими методами статистичної обробки, не характеризує розсіювання результатів вимірювань. Для обробки даних, які є членами несиметричних і «забруднених» (з обтяженими хвостами) розподілів, розроблені методи, що дозволяють отримувати надійніші та стійкіші до викидів оцінки результатів вимірювань. ***Стійкі (робастні) методи дослідження***, певною мірою, є альтернативою методам виключення грубих помилок, що здійснюються під час попередньої обробки результатів. Виключення із загальної сукупності значень, що різко виділяються, може призвести до видалення великої кількості результатів досліджень, тому результат не відповідатиме дійсності. Ця проблема актуальна при невеликих розмірах вибірки. *Робастні методи оцінювання враховують наявність грубих помилок і дозволяють при цьому досить надійно оцінювати параметри, що характеризують результат*.

Створення стійких методів оцінювання було викликане прагненням поліпшити існуючі схеми методу найменших квадратів так, щоб викиди якнайменше впливали на кінцеві результати оцінювання. Адже відомо, що оцінки методу найменших квадратів ефективні, якщо похибки вимірювань розподілені за нормальним законом.

4.2.1 Метод Пуанкаре

За ***методом Пуанкаре*** результати досліджень розміщують у вигляді варіаційного ряду і, відкидаючи по k крайніх членів з кожного кінця ряду, отримують скорочену вибірку та ***скорочене середнє*** за формулою

$$T(\alpha)=\frac{1}{N-2k}\sum_{i=k+1}^{N-k}x_i\ , \tag{4.9}$$

де N – розмір вибірки;

k – число відкинутих значень, $k \le \alpha N$ – ціла частина від добутку αN;

$\alpha(\xi)$ – деяка функція засмічення вибірки (ξ – відношення кількості підозрілих даних до розміру вихідної вибірки), $0 \le \alpha \le 0.5$ (див. табл.4.5).

Таблиця 6.5 – Значення α для розрахунку стійких оцінок Пуанкаре *T(α)* та Вінзора *W(α)*

ξ	0	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0,15
α	0	0.004	0.008	0.015	0.026	0.043	0.081	0.127	0.164
ξ	0.2	0.25	0.3	0.4	0.5	0.65	0.8	1	
α	0.194	0.222	0.247	0.291	0.332	0.386	0.436	0.5	

При $\alpha \to 0.5$ скорочене середнє наближається до медіани вибірки (***медіана*** – число, яке є серединою множини чисел: половина чисел мають значення більші за медіану, а половина – менші):

– $med = x_{k+1}$ при непарному N;

– $med = (x_k + x_{k+1})$ при парному N.

При $\alpha \to 0$ виходить звичайне середнє $\bar{x}$.

Якщо помилкові дані розміщені у верхній (або нижній) частині вибірки, то ці дані видаляються із сукупності. Однак для того, щоб видалення суттєво не вплинуло на зміну розрахованого значення відносно істинного, з нижньої (або, відповідно, верхньої) частини остаточної сукупності видаляються k перших значень. Скорочене середнє визначається формулою (4.9).

Якщо ж помилкові дані розташовані з обох кінців сукупності, вона модифікується таким чином, щоб мінімізувати кількість вірних даних, які

будуть видалені з вибірки. Тобто з одного кінця впорядкованого ряду видаляються всі помилкові дані, а з іншого – того, де їх було менше, крім помилкових, видаляються ще й вірні дані.

Метод Пуанкаре є досить простим, але має недолік – значно скорочує остаточну послідовність даних.

Наприклад, розглянемо вибірку: 21.2, 19.9, 24.1, 20.4, 21.5, 19.7, 18.9, 23.6, 19.7, 20.6. У нашій вибірці два значення дуже відрізняються від інших – це 24.1 та 23.6. Відхилення особливо помітні, якщо даний ряд упорядкувати за зростанням: 18.9, 19.7, 19.7, 19.9, 20.4, 20.6, 21.2, 21.5, 23.6, 24.1.

Для даних вибірки розрахуємо середнє за Пуанкаре. Так ξ=0.2 (ділимо кількість підозрілих даних (2) на кількість даних всієї вибірки (10)), α =0.194 (з табл.4.5), k=2.

Оскільки була визначена наявність двох підозрілих значень у верхній частині ряду, то для обчислення залишаються 6 значень. Тоді за формулою (4.9) скорочене середнє дорівнює

$$T = \frac{1}{(10-2\cdot 2)}\sum_{i=2+1}^{10-2} x_i = \frac{1}{6}\sum_{i=3}^{8} x_i = \frac{19.7+19.9+20.4+20.6+21.2+21.5}{6} = 20.55$$

і близьке до медіани вибірки (20.50), що є перевіркою правильності відкидання.

4.2.2 Метод Вінзора

До скорочених середніх близькі також вінзоровані середні. В методі Вінзора крайні (підозрілі) елементи вибірки не відкидають, а замінюють на найближчі до них із залишених членів ряду або на зважені з деякими вагами (вінзоровані середні).

За ***методом Вінзора*** середнє (***вінзороване середнє***) визначають за формулою

$$W(\alpha) = \frac{1}{N}\left(\sum_{i=k+1}^{N-k} x_i + k\left(x_{k+1} + x_{N-k}\right)\right). \qquad (4.10)$$

Оцінка (4.10) є модифікацією скороченого середнього (4.9). Але, на відміну від скороченої середньої оцінки, в оцінці Вінзора 2k крайніх елементів вибірки не відкидаються, а k лівих крайніх замінюються на елемент x_{k+1}, а k правих крайніх замінюються на елемент x_{N-k} (тобто замінюються на найближчі до них) або замінюються вінзорованими середніми.

Розглянемо застосування метода Вінзора для вибірки з попереднього прикладу. Відсортувавши вибірку за зростанням (18.9, 19.7, 19.7, 19.9, 20.4, 20.6, 21.2, 21.5, 23.6, 24.1) і виявивши два значення (k=2), які дуже відрізняються від інших (24.1 та 23.6), отримуємо: ξ=0.2, α=0.194 (з табл.4.5). Тоді $W = \frac{1}{10}\left(\sum_{i=2+1}^{10-2} x_i + 2\left(x_{2+1} + x_{10-2}\right)\right) = 0.1 \cdot \left(\sum_{i=3}^{8} x_i + 2\left(x_3 + x_8\right)\right) =$

$= 0.1 \cdot \left(123.3 + 2\left(19.7 + 21.5\right)\right) = 20.57$ і близьке до медіани вибірки (20.50), а наша вибірка набуде такого вигляду: 20.57, 20.57, 19.7, 19.9, 20.4, 20.6, 21.2, 21.5, 20.57, 20.57.

Якщо ж підозрілі елементи вибірки замінити на найближчі до них, отримаємо таку вибірку: 19.7, 19.7, 19.7, 19.9, 20.4, 20.6, 21.2, 21.5, 21.5, 21.5.

4.2.3 Метод Хьюбера

Окрім розглянутих методів оцінювання, широкого застосування набув класичний підхід Хьюбера. При цьому використовується деяка величина k, яка розраховується з урахуванням міри засмічення статистичної сукупності та визначає крок модифікації даних, що сильно відрізняються. За ***методом Хьюбера*** середнє обчислюють формулою

$$\hat{\theta} = \frac{1}{N - \left(n_1 + n_2\right)}\left(\sum_{|x_i - \theta| < k} x_i + k\left(n_1 + n_2\right)\right), \qquad (4.11)$$

де $\hat{\theta}$ – стійка оцінка, що визначається за допомогою ітеративних процедур;

k – величина, яка допускається як відхилення від центру сукупності та приймає постійні значення з урахуванням питомої ваги грубих помилок у сукупності даних (див. табл.4.6);

n_1, n_2 – кількість вимірювань x_i, що потрапляють всередину інтервалу $(-\infty;\theta-k)$ і $(\theta+k;+\infty)$, відповідно;

θ – початкова оцінка.

Таблиця 4.6 – Значення k для розрахунку стійкої оцінки Хьюбера $\hat{\theta}$

ξ	0	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0,15
k	0	2.63	2.435	2.16	1.945	1.717	1.399	1.14	0.98
ξ	0.2	0.25	0.3	0.4	0.5	0.65	0.8	1	
k	0.862	0.766	0.685	0.55	0.436	0.291	0.162	0	

При розрахунках за формулою (4.11) початковою оцінкою θ можуть бути середнє арифметичне $\bar{x}$ або медіана, визначена за вибіркою. Потім на кожній ітерації дані вибірки розділяють на три частини. До однієї частини потрапляють «істинні» ознакові значення, які залишаються без зміни ($|x_i-\theta|<k$). До двох інших частин сукупності (для $x_i>\theta+k$ та $x_i<\theta-k$) – «помилки». Вони не виключаються з розгляду, а замінюються, відповідно, на величини x_i-k та x_i+k. За «істинними» та модифікованими даними кожного разу визначається нова оцінка $\hat{\theta}$ – і ітерація повторюється. *Ітерації повторюються доти, поки всі дані не опиняться в інтервалі «істинних» значень:* $|x_i-\hat{\theta}|<k$.

Оцінка за методом Хьюбера є ефективною, але швидко втрачає оптимальність зі збільшенням міри засмічення вибірки (збільшенням ξ).

Розглянемо застосування метода Хьюбера для вибірки з останнього прикладу. Розділивши кількість помилок (2) на кількість даних вибірки (10) (ξ=0.2), з табл.4.6 отримаємо k =0.862. Знайдемо початкову оцінку як середнє значення: $\theta=\bar{x}=20.96$. Далі розіб'ємо сукупність даних на три групи: ті, що незначно відрізняються від θ; суттєво менші за величину θ та суттєво

більші за θ. Потім відповідним чином модифікуємо x_i, якщо $x_i > \theta + k$ або $x_i < \theta - k$. Після цього за формулою (4.11) розрахуємо нову оцінку $\hat{\theta}_1$ з урахуванням даних, модифікованих на попередній ітерації, де: n_1=1 (кількість вимірювань x_i, що потрапляють всередину інтервалу $(-\infty;20.098)$, тобто кількість елементів групи III); n_2=2 (кількість вимірювань x_i, що потрапляють всередину інтервалу $(21.822;+\infty)$, тобто кількість елементів групи II); $\sum_{|x_i-\theta|<k} x_i = 145.568$ – сума всіх x_i, що потрапили до групи I (після перерозподілу даних): $\hat{\theta}_1 = \frac{1}{10-(1+2)} \cdot (145.586 + 0.862 \cdot (1+2)) \approx 21.167$.

Повторюємо наші дії доти, поки не виконуватиметься рівність $|x_i - \hat{\theta}_i| < k$ для всіх x_i чергової ітерації.

Так, $\hat{\theta}_2 = \frac{1}{10-(0+1)} \cdot (188.086 + 0.862 \cdot (0+1)) \approx 20.994$.

Результати розрахунків кожної ітерації наведені в табл.4.7. В таблиці жирним курсивом виділено дані II та III груп, які після проведеної модифікації потрапляють до групи I «істинних» значень (див. відповідний черговий перерозподіл даних з груп).

Таблиця 4.7 – Результати розрахунків методом Хьюбера

Ітерація 0			
Вихідні значення: 18.9000, 19.7000, 19.7000, 19.9000, 20.4000, 20.6000, 21.2000, 21.5000, 23.6000, 24.1000, $\theta = \bar{x} = 20.96, k = 0.862$	**I** $\lvert x_i - \theta \rvert < k$	**II** $x_i > \theta + k = 21.822$	**III** $x_i < \theta - k = 20.098$
	20.400	23.600	18.900
	20.600	24.100	19.700
	21.200		19.700
	21.500		19.900
Модифікація вихідних значень II та III класів	*без зміни*	$x_i - k$	$x_i + k$
	20.400	22.738	19.762
	20.600	23.238	***20.562***
	21.200		***20.562***
	21.500		***20.762***

Продовження Таблиці 4.7

Перерозподіл даних з груп	$\lvert x_i - \theta \rvert < k$	$x_i > \theta + k = 21.822$	$x_i < \theta - k = 20.098$
	20.400	22.738	19.762
	20.562	23.238	
	20.562		
	20.600		
	20.762		
	21.200		
	21.500		
Ітерація 1			
Нові значення: 19.7620, 20.4000, 20.5620, 20.5620, 20.6000, 20.7620, 21.2000, 21.5000, 22.7380, 23.2380, $\hat{\theta}_1 = 21.167$	$\lvert x_i - \hat{\theta}_1 \rvert < k$	$x_i > \hat{\theta}_1 + k = 22.029$	$x_i < \hat{\theta}_1 - k = 20.305$
	20.400	22.738	19.762
	20.562	23.238	
	20.562		
	20.600		
	20.762		
	21.200		
	21.500		
Модифікація нових значень з груп II та III	*без зміни*	$x_i - k$	$x_i + k$
	20.400	***21.876***	***20.624***
	20.562	22.376	
	20.562		
	20.600		
	20.762		
	21.200		
	21.500		
Перерозподіл даних з груп	$\lvert x_i - \hat{\theta}_1 \rvert < k$	$x_i > \hat{\theta}_1 + k = 22.029$	$x_i < \hat{\theta}_1 - k = 20.305$
	20.400	22.376	–
	20.562		
	20.562		
	20.600		
	20.624		
	20.762		
	21.200		
	21.500		
	21.876		

Продовження Таблиці 4.7

Ітерація 2			
Нові значення: 20.4000, 20.5620, 20.5620, 20.6000, 20.6240, 20.7620, 21.2000, 21.5000, 21.8760, 22.3760 $\hat{\theta}_2 = 20.994$	$\lvert x_i - \hat{\theta}_2 \rvert < k$	$x_i > \hat{\theta}_2 + k = 21.856$	$x_i < \hat{\theta}_2 - k = 20.132$
	20.400	21.876	–
	20.562	22.376	
	20.562		
	20.600		
	20.624		
	20.762		
	21.200		
	21.500		
Модифікація нових значень з груп II та III	*без зміни*	$x_i - k$	$x_i + k$
	20.400	***21.014***	–
	20.562	***21.514***	
	20.562		
	20.600		
	20.624		
	20.762		
	21.200		
	21.500		
Перерозподіл даних з груп	$\lvert x_i - \hat{\theta}_2 \rvert < k$	$x_i > \hat{\theta}_2 + k = 21.856$	$x_i < \hat{\theta}_2 - k = 20.132$
	20.400	–	–
	20.562		
	20.562		
	20.600		
	20.624		
	20.762		
	21.014		
	21.200		
	21.500		
	21.514		

Як видно з таблиці, рівність $\lvert x_i - \hat{\theta}_2 \rvert < k$ виконується для всіх x_i останньої (2-ої) ітерації. Тому остаточна вибірка матиме вигляд: 20.400, 20.562, 20.562, 20.600, 20.624, 20.762, 21.014, 21.200, 21.500, 21.514, 21.514.

Контрольні питання

1. Поняття грубої помилки (промаху, викиду).
2. Довірча імовірність та довірчий інтервал.
3. Критерії виключення грубих помилок.
4. Оптимізація засмічення вибірки за допомогою робастних методів.