Лекція 4. ДОДАТКОВІ СТАТИСТИЧНІ ОЦІНКИ ПАРАМЕТРІВ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ

4.1. Інші оцінки міри варіативності

Квантилем називають значення ранжированої змінної, що відокремлює від варіаційного ряду певну частку обсягу сукупності. Це значення, яке випадкова величина не перевищує із заданою ймовірність.

В математичній статистиці використовуються такі квантилі:

- процентилі $(P_1, P_2, \ldots, P_{99})$;
- децилі $(D_1, D_2, \ldots, D_9)$;
- квінтилі (K_1, K_2, K_3, K_4);
- квартилі (Q_1, Q_2, Q_3).

Процентилі ділять упорядковану сукупність на сто частин, тобто відокремлюють від сукупності по 0,01 частині (по 1%).

Квартилі ділять сукупність на чотири частини. Перший квартиль Q_1 відокремлює зліва 0,25 обсягу сукупності. Другий квартиль Q_2 ділить сукупність на дві рівні за обсягом частини (по 0,5), він називається медіаною. Нарешті, третій квартиль Q_3 відокремлює зліва 0,75 обсягу сукупності (рис. 4.1).

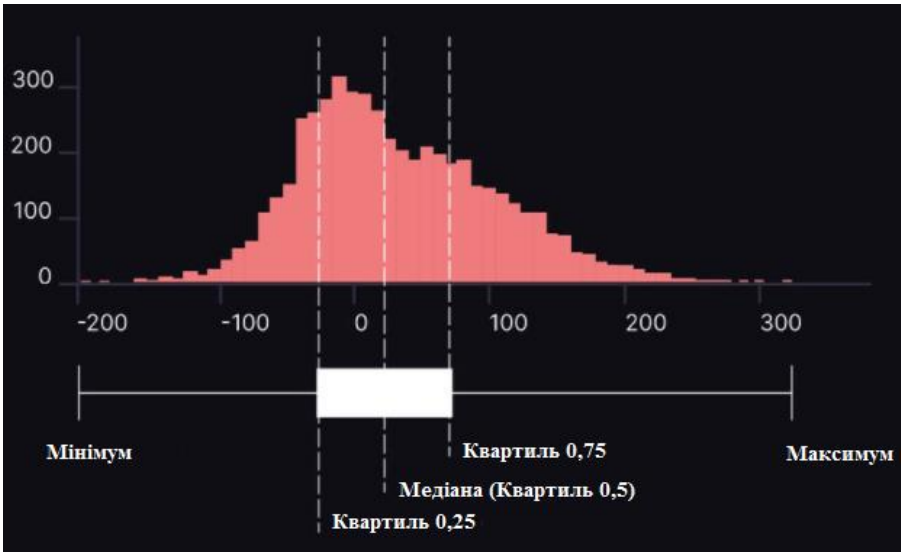

Рис. 4.1. Графічне зображення значень квартиль

Різниця 75% та 25% рівнями значень варіанти називається *міжквартильним розмахом*.

$$IQR = Q_3 - Q_1.$$

Міжквартильний розмах більш стійкий до викидів.

Діаграми розмаху (рис.4.2) можуть зображатися горизонтально або вертикально. Розмір «коробки» відповідає відстані між нижнім і верхнім квартилями, тобто міжквартильним розмахом.

Відстань між кінцями «вусів» відповідає усьому діапазону даних, які спостерігалися в експерименті (від мінімума до максимума).

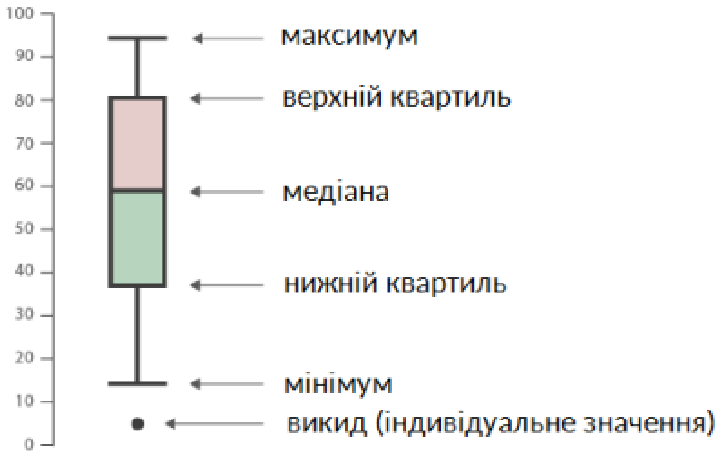

Рис. 4.2. Графічне зображення значень box-and-whiskers plot

В деяких випадках статистичні програми можуть розцінювати певні значення, що сильно вибиваються з основного масиву даних, як викиди (outliers) та грубі помилки. Вони можуть позначатися окремими точками, що лежать за межами «вусів». Виявлення та аналіз причин появи таких значень є важливим етапом попереднього аналізу даних, тому що наявність навіть одного такого значення може сильно спотворити значення багатьох статистик.

Нормально розподілені дані часто зображаються у формі стовпчиків, висота яких позначає середнє значення, з «вусами», розмах яких дорівнює $\pm\sigma$. Стовпчики діаграм можуть не зображатися, тоді середнє значення позначатиметься точкою.

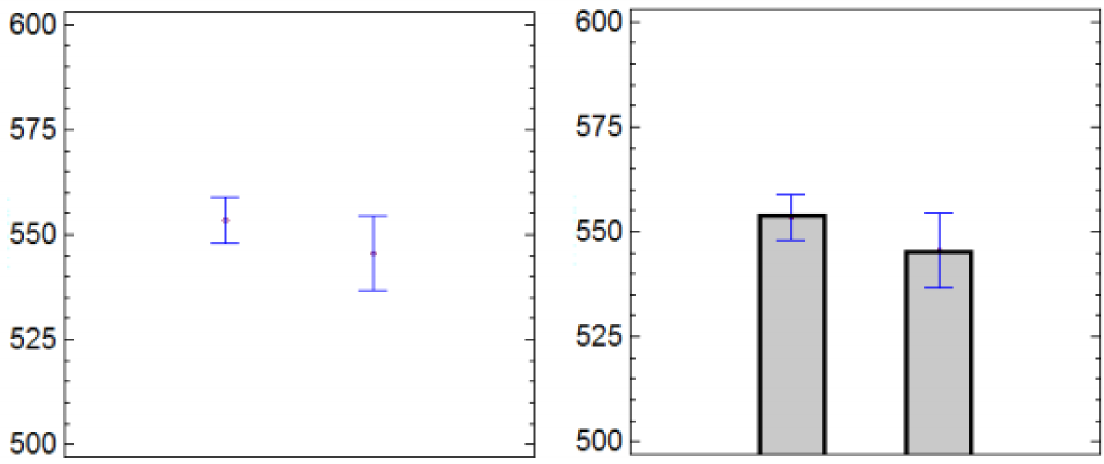

Рис. 4.3. Два варіанти графічного представлення одного і того ж набору нормально розподілених даних

Між різницями квантилями існують певні співвідношення, наприклад, між квартилями і процентилями такі: $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$ (рис. 4.4)

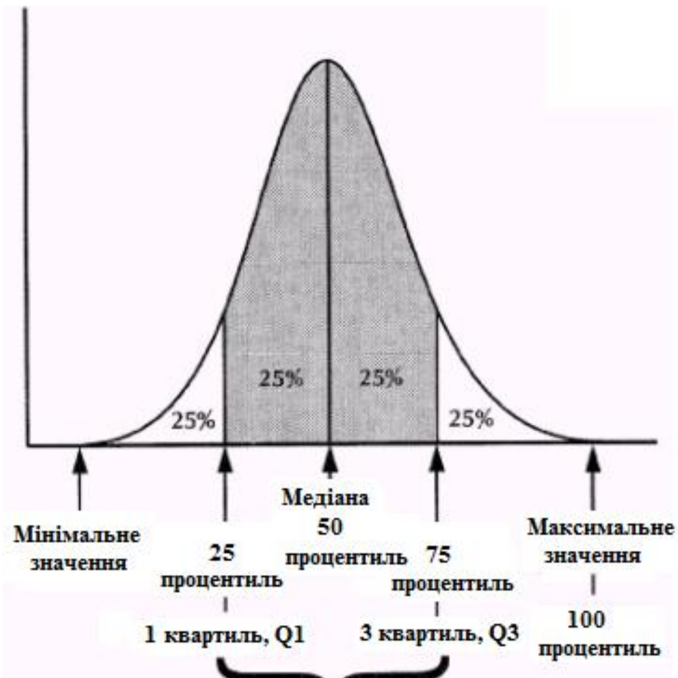

Рис. 4.4. Функція Гаусса

Обчислювати квантилі можна графічно або за таблицями. Так видно, що 25-й процентиль P_{25} і 1-й квартиль Q_1 (рис. 4.5) дорівнюють значенню 3 ($P_{25} = 3$ і $Q_1 = 3$). Отже, нижче за це значення знаходяться 25% усіх значень.

Аналогічно можна знайти інші співвідношення, наприклад P_{75} і Q_3 (75-й процентиль і 3-й квартиль) дорівнюють 5. Нижче за це значення знаходяться 75% всіх значень.

	A	B	C	D	E	F	G
1	Вибірка			Розрахунки			
2	i	Первинні дані	Упорядковані дані	Номери квантилів		Значення квантилів	
3				Процентиль	Квартиль	Процентиль	Квартиль
4	1	7	**1**	0,00	0	**1**	**1**
5	2	3	2	0,05		**2**	
6	3	6	2	0,10		**2**	
7	4	4	3	0,15		**3**	
8	5	5	3	0,20		**3**	
9	6	6	**3**	0,25	1	**3**	**3**
10	7	4	4	0,30		**4**	
11	8	6	4	0,35		**4**	
12	9	4	4	0,40		**4**	
13	10	1	4	0,45		**4**	
14	11	5	**4**	0,50	2	**4**	**4**
15	12	6	5	0,55		**5**	
16	13	3	5	0,60		**5**	
17	14	4	5	0,65		**5**	
18	15	5	5	0,70		**5**	
19	16	2	**5**	0,75	3	**5**	**5**
20	17	3	6	0,80		**6**	
21	18	2	6	0,85		**6**	
22	19	4	6	0,90		**6**	
23	20	5	6	0,95		**6**	
24		5	**7**	1,00	4	**7**	**7**

Рис. 4.5. Співвідношення квантилів

Основні формули для обчислення квартилів та децілий:

$$Q_1 = x_{Q_1} + h \cdot \frac{\frac{n}{4} - S_{Q_1 - 1}}{n_{Q_1}} \text{ - перший квартиль;} \tag{4.1}$$

$$Q_3 = x_{Q_3} + h \cdot \frac{\frac{3}{4} \cdot n - S_{Q_3 - 1}}{n_{Q_3}} \text{ - третій квартиль;} \tag{4.2}$$

$$D = x_D + h \cdot \frac{0{,}1 \cdot n - S_{D-1}}{n_D} \text{ - деціли,} \tag{4.3}$$

де, $x_{Q_1}\left(x_{Q_3}\right)$ та x_D - нижня межа інтервалу першого (третього) квартилів та децильного інтервалу; h - розмір (крок) інтервалу; $n_{Q_1}\left(n_{Q_3}\right)$ - частота у варіант, які потрапляють у межу $x_{Q_1}\left(x_{Q_3}\right)$ відповідно; n_D - частоти відповідних децильних інтервалів; $S_{Q_1-1}\left(S_{Q_3-1}\right)$ та S_{D-1} - сума накопичених частот інтервалів, що передують відповідним квартильним $S_{Q_1}\left(S_{Q_3}\right)$ та децільним відповідно; n - обсяг вибірки.

Приклад 4.1. Інтегральний статистичний розподіл вибірки 200 корів за живою масою задано у вигляді таблиці:

Маса, кг	400-420	420-440	440-460	460-480	480-500	500-520	520-540
К-сть	5	19	38	63	44	26	5

Знайти квартилі та децили. Зробити відповідні висновки.

Розв'язання:

Перший квартиль становить:

$$Q_1 = x_{Q_1} + h\cdot\frac{\frac{n}{4} - S_{Q_1-1}}{n_{Q_1}} = 440 + 20\cdot\frac{\frac{200}{4} - 24}{38} = 453{,}7 \text{ кг.}$$

Це означає, що одна чверть корів має живу масу 453,7 кг і менше, а три чверті –більш як 453,7 кг.

Третій квартиль становить:

$$Q_3 = x_{Q_3} + h\cdot\frac{\frac{3}{4}\cdot n - S_{Q_3-1}}{n_{Q_3}} = 480 + 20\cdot\frac{\frac{3}{4}\cdot 200 - 125}{44} = 491{,}4 \text{ кг.}$$

Це означає, що три чверті корів мають живу масу до 491,4 кг, а одна чверть –більш як 491,4 кг.

Децили:

$$D = x_D + h\cdot\frac{0{,}1\cdot n - S_{D-1}}{n_D} = 420 + 20\cdot\frac{0{,}1\cdot 200 - 5}{19} = 435{,}8 \text{ кг.}$$

Отже, десята частина всіх корів має живу масу 435,8 кг і менше, а решта (90%) – більш як 435,8 кг.

4.2. Статистичний контроль якості

Контрольна карта середніх арифметичних технологічного процесу.

Якщо генеральна сукупність має нормальний (або близький до нормального) розподіл з середнім (номінальним) значенням $T_{ном}$ і середнім відхиленням σ, вибірковий розподіл вибіркової середньої також є нормальним і має таке саме середнє значення та стандартну помилку, рівну $\frac{\sigma}{\sqrt{n}}$, де n - об’єм вибірки.

Для будь-якого нормального розподілу між граничними значеннями, рівними $\left(T_{ном} \pm \frac{2\sigma}{\sqrt{n}}\right)$ стандартному відхиленню, знаходиться приблизно 95% розподілу. Ймовірність того, що отримане значення виявиться більшим, ніж стандартне відхилення $\left(T_{ном} + \frac{2\sigma}{\sqrt{n}}\right)$, становить 2,5%, або один із 40, ймовірність отримання значення, меншого стандартного відхилення $\left(T_{ном} - \frac{2\sigma}{\sqrt{n}}\right)$ також становить 2,5%.

Аналогічно інтервал $\left(T_{ном} \pm \frac{3\sigma}{\sqrt{n}}\right)$ стандартне відхилення охоплює близько 99,8% розподілу. Ймовірність того, що отримане значення перевищує стандартне відхилення $\left(T_{ном} + \frac{3\sigma}{\sqrt{n}}\right)$ або виявиться меншим, ніж стандартного відхилення $\left(T_{ном} - \frac{3\sigma}{\sqrt{n}}\right)$, становить 0,1%, тобто ця подія матиме місце в одному випадку із тисячі.

Графічна ілюстрація цих крайніх значень показана на рис. 4.6. вибіркового розподілу вибіркової середньої арифметичного технологічного процесу.

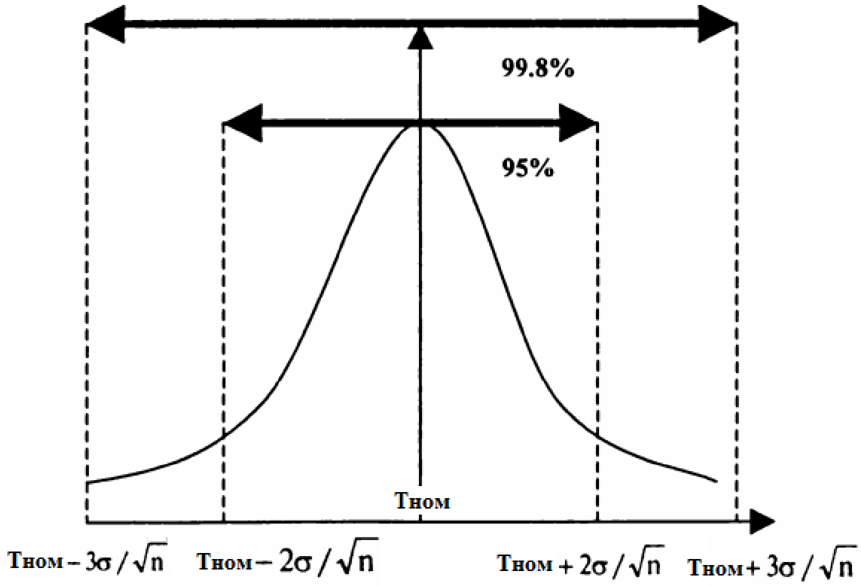

Рис. 4.6.

Для побудови цього графіка необхідно, щоб значення $T_{ном}$ та дисперсії були відомі. 95%-ві межі розподілу називаються верхньою і нижньою *попереджувальними межами*. 98%-ві межі розподілу називають верхньою і нижньою *межами регулювання*.

Побудова контрольної карти полягає в нанесенні на графік вибіркових середніх у відповідності з номером вибірки (рис. 4.7).

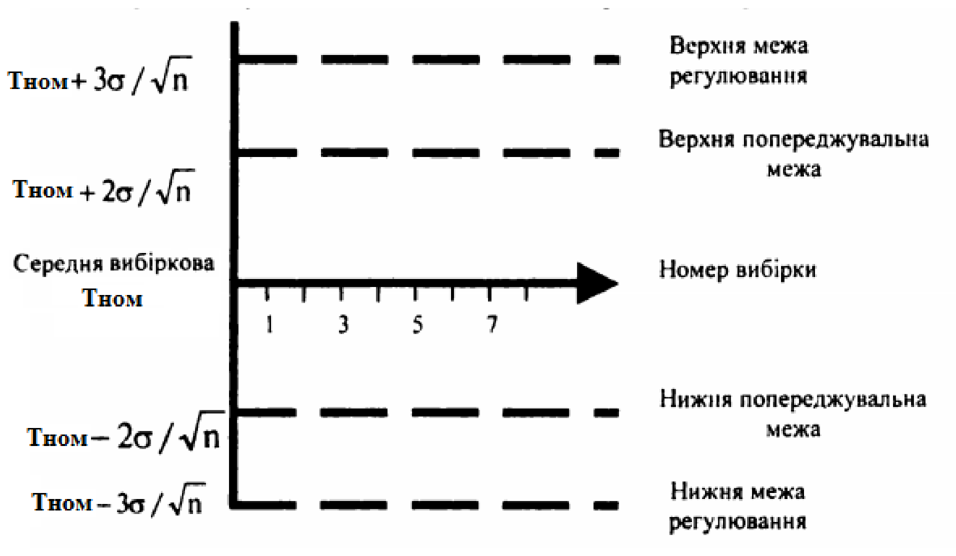

Рис. 4.7.

Стандартна процедура використання контрольних карт складається з таких кроків:

1. Через рівні проміжки часу проводиться вибірка об’ємом *n* і розраховується вибіркова середня.

2. Отримані значення вибіркових середніх наносяться на контрольну карту відповідно до номера вибірки.

3. Якщо вибіркова середня лежить за межами регулювання, технологічний процес зупиняється з метою виявлення невипадкових причин варіації.

4. Якщо два послідовно отриманих значення вибіркових середніх знаходяться на проміжку між попереджувальною межею і межею регулювання, проводяться невідкладні дії із зупинки процесу виробництва та виявлення неполадок. Якщо деяке середнє занесення до карти лежить за попереджувальними межами, наступна вибірка проводиться відразу ж, до моменту проведення чергової вибірки.

5. Якщо точки на графіку утворюють явний зростаючий чи спадаючий тренд, вживаються певні заходи навіть у випадках, коли ці точки знаходяться в попереджувальних межах. Цей тренд може виявитись індикатором наявності невипадкових причин, наприклад, зниження параметрів наладки верстата.

Приклад 4.2. Проводиться фасування чаю в упаковки об’ємом по 125гр. Через кожні півгодини проводиться випадкова вибірка об’ємом 5 упаковок. Кожну упаковку зважують. Результати шести послідовних вибірок наведені у таблиці:

Номер вибірки	1	2	3	4	5	6
Вага упакрвки, гр	125,1	124,9	125,2	125,0	124,8	124,9
	125,3	125,0	125,1	125,0	124,8	125,1
	125,1	125,1	125,3	124,7	125,2	125,0
	124,8	124,9	125,0	125,2	125,1	124,9
	125,1	124,7	125,1	125,1	124,9	125,2

Побудувати за цими даними контрольну карту середнього арифметичного та описати функціонування процесу розфасовки.

Розв'язання. Центральна вісь контрольної карти відповідає рівню $T_{ном} = 125$ гр.

Попереджувальні межі будуються на рівнях:

$T_{ном} \pm 2 \cdot \frac{\sigma}{\sqrt{n}} = 125 \pm 2 \cdot \frac{0,15}{\sqrt{5}} = 125 \pm 0,134$, тобто для 124,866 та 125,134 гр.

Межі регулювання будуються на рівнях:

$T_{ном} \pm 3 \cdot \frac{\sigma}{\sqrt{n}} = 125 \pm 3 \cdot \frac{0,15}{\sqrt{5}} = 125 \pm 0,201$, тобто для 124,799 та 125,201 гр.

Номер вибірки	1	2	3	4	5	6
Середнє значення, $\overline{x}_B$ гр	125,08	124,92	125,14	125,0	124,96	125,02

Нанесемо середні значення на контрольну карту (рис. 4.8).

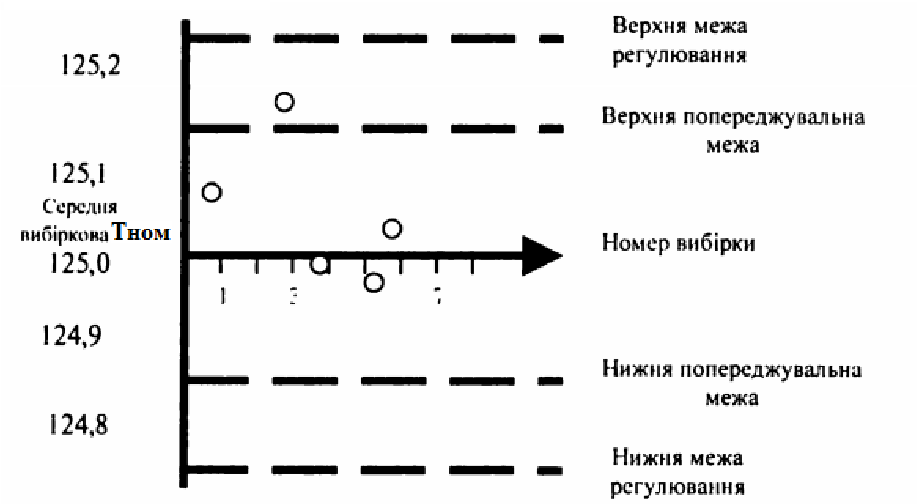

Рис. 4.8.

Середнє значення вибірки 3 (125,14) знаходиться вище верхньої попереджувальної межі, однак середнє значення наступної вибірки знаходиться в середині контрольних границь. Тобто можна запропонувати, що приводів для хвилювання немає. Припускається, що вибірка 4 проводиться відразу після вибірки 3, в якій були виявлені деякі відхилення параметрів.