



# МОДЕЛІ ТА МЕТОДИ ТЕОРІЇ СИСТЕМ І СИСТЕМНОГО АНАЛІЗУ

УДК 519.2:621.317

## ДОСЛІДЖЕННЯ ВЛАСТИВОСТЕЙ ОЦІНОК ДИСПЕРСІЇ, ОТРИМАНИХ ЗА ГРУПОВИМ МЕТОДОМ

Бахрушин В.Є.

*Гуманітарний університет "Запорізький інститут державного та муніципального управління",**вул. Жуковського, 70-б, Запоріжжя, Україна, 69002**Vladimir.Bakhrushin@zhu.edu.ua*

### Вступ

Дисперсія є однією з основних характеристик вибірок і характеризує відхилення елементів сукупності від середнього. При побудові описової статистики використовують різні методи її оцінювання. Для ознак, що визначаються у кількісних шкалах, найчастіше вибіркочну дисперсію розраховують [1; 2] за формулою:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sum_{i=1}^n x_i^2 - n \left( \sum_{i=1}^n x_i \right)^2}{n(n-1)}. \quad (1)$$

де  $x_i$  – значення ознаки для  $i$ -го об'єкта,  $\bar{x}$  – вибіркоче середнє,  $n$  – кількість об'єктів.

Стандартне відхилення дисперсії у цьому разі визначають за формулою  $s_{\sigma^2} = \sigma^2 \sqrt{\frac{2}{n}}$ .

Якщо вихідні дані задано у вигляді частот розподілу, тобто вихідні дані згруповані за певними класовими інтервалами, дисперсію можна оцінити за формулою [1]:

$$\sigma^2 = \frac{1}{n-1} \left[ \sum_{i=1}^k v_i b_i^2 - \frac{1}{n} \left( \sum_{i=1}^k v_i b_i \right)^2 \right], \quad (2)$$

де  $b_i$  ( $i = 1, 2, \dots, k$ ) – середини класових інтервалів,  $v_i$  – частоти,  $k$  – кількість класових інтервалів.

Часто зазначають, що ця формула дає завищену оцінку дисперсії [2; 3]. Для її корегування вводять поправку Шеппарда і визначають уточнене значення за формулою:

$$s'^2 = \sigma^2 - \frac{h^2}{12}, \quad (3)$$

де  $h$  – інтервал між групами, який за рівних відстаней між групами збігається з величиною класового інтервалу.

Але практика її застосування свідчить, що значення дисперсії, розраховані за формулою (2) в окремих випадках можуть бути меншими, ніж значення, що визначають для тих самих даних за формулою (1).



Умови, за яких формула (2) є обґрунтованою, чітко не визначені. Найчастіше рекомендують використовувати поправку Шеппарда, якщо обсяг вибірки перевищує 500 елементів [2]. У фундаментальному довіднику [3] зазначають, що застосування цієї поправки є правомірним, якщо для теоретичного розподілу обидва хвости мають високий порядок зіткнення з віссю абсцис.

Крім того, згідно з наявними в літературі даними, на результати, одержувані за згрупованими даними, може істотно впливати кількість інтервалів розбиття [4; 5]. Існує багато рекомендацій з вибору цієї величини [6], але вони є досить суперечливими й недостатньо обґрунтованими теоретично.

**Метою** даної роботи було дослідження статистичних властивостей оцінок дисперсії, що розраховуються за формулою (2).

### Методика дослідження

Аналізовані послідовності генерували за допомогою відповідної процедури пакета аналізу електронних таблиць MS Excel. При цьому використовували нормально розподілені послідовності обсягом від 200 до 30000 елементів із середніми значеннями 1–10 і стандартним відхиленням 1–50. Для кожного набору значень середнього, стандартного відхилення й обсягу вибірки усі досліджувані послідовності генерували разом, а потім проводили оцінювання їх параметрів. Кількість згенерованих послідовностей для всіх наборів параметрів коливалася від 100 до 300.

Дисперсії вибірок  $\sigma_1$  та  $\sigma_2$  розраховували за формулами (1) і (2) відповідно. В останньому разі вихідну вибірку  $x_i$  попередньо розбивали на визначену кількість  $k$  класових інтервалів. Ліву межу першого інтервалу  $x_0$  визначали як  $x_{\min} - 0,001|x_{\min}|$ , а праву межу останнього інтервалу  $x_{20}$  – за формулою:  $x_{\max} + 0,001x_{\max}$ , де  $x_{\min}$  та  $x_{\max}$  – мінімальне й максимальне значення аналізованої вибірки. Ширину класових інтервалів розраховували за формулою:

$$h = \frac{x_{20} - x_0}{20}$$

Інші межі інтервалів послідовно розраховували за формулою:  $x_i = x_{i-1} + h$ . Частоти для інтервалів розраховували за допомогою функції "ЧАСТОТА". При цьому задавали такі параметри: масив даних – посилання на адреси комірок, де знаходилися елементи аналізованої послідовності; масив інтервалів – посилання на адреси комірок, де було вказано праві межі інтервалів. Кількість класових інтервалів коливалася від 11 до 20.

Потім розраховували відношення  $\alpha = \sigma_2 / \sigma_1$ . Для масиву значень  $\alpha$ , що відповідали всім отриманим послідовностям, за допомогою вбудованих статистичних функцій електронних таблиць MS Excel визначали середнє значення, стандартне відхилення, мінімальне й максимальне значення, а також будували функції розподілу отриманих оцінок дисперсії. Потім аналізували вплив обсягу вихідних вибірок, їх середнього значення й стандартного відхилення на вказані параметри.

Проведений аналіз свідчить, що із зростанням кількості інтервалів графік функції розподілу зміщується вліво, а середнє значення одержуваних оцінок дисперсії зменшується. При цьому з усіх проаналізованих випадків лише для послідовностей, що містили 200 елементів, при розбитті на 20 інтервалів отримали середнє значення дисперсії, що було меншим ніж значення, розраховане за формулою (1). Тому наведені нижче дані відповідають випадкам, коли використовували розбиття на 20 класових інтервалів.

### Результати дослідження

Отримані у роботі результати свідчать про те, що найбільш істотний вплив на досліджувані параметри виявляє кількість елементів послідовності. Водночас вплив на них середнього значення й стандартного відхилення вихідної послідовності є значно меншим.



Для вибірок малого обсягу функція розподілу параметра  $\alpha$  не залежить від стандартного відхилення (рис. 1), а також за даними графічного тесту (рис. 2) і значеннями критерію Стьюдента відповідає нормальному закону розподілу.

При збільшенні обсягу генерованих вибірок розкид між функціями розподілу параметра  $\alpha$ , що відповідають вибіркам з різними значеннями стандартного відхилення, залишається несуттєвим. Але при обсязі вибірок 5000 і 30000 виявляються істотні відхилення функцій розподілу від нормального закону (рис. 3, 4).

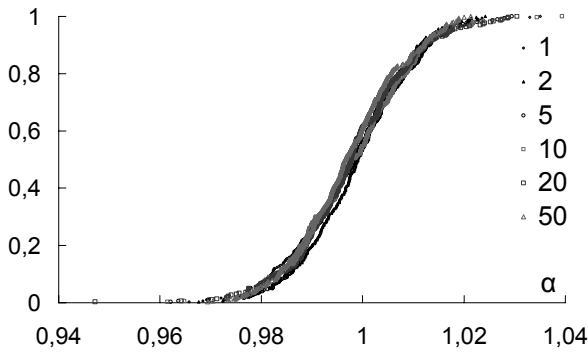


Рис. 1 Емпіричні функції розподілу для  $n = 200$  і середнього значення 10

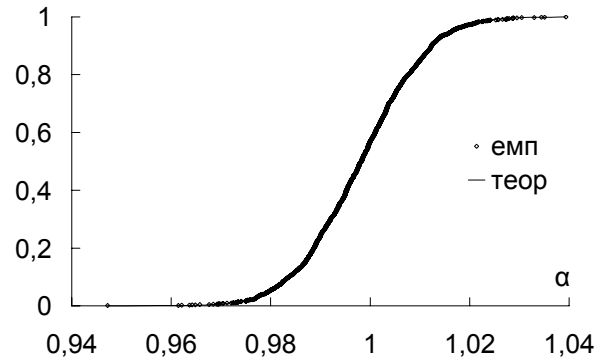


Рис. 2. Емпірична й теоретична функції розподілу для  $n = 200$

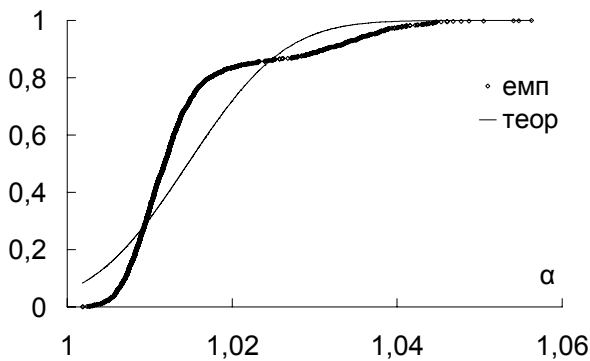


Рис. 3. Емпірична й теоретична функції розподілу для  $n = 5000$

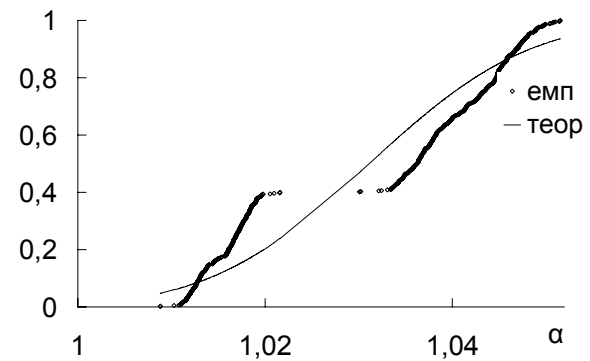


Рис. 4. Емпірична й теоретична функції розподілу для  $n = 30000$

На рис. 5–8 показано вплив кількості елементів ( $n$ ), середнього значення ( $\langle x \rangle$ ) та стандартного відхилення ( $\sigma_x$ ) вихідних послідовностей на середнє значення, стандартне відхилення, мінімальне й максимальне значення масивів, створених із значень параметрів  $\alpha = \sigma_2 / \sigma_1$ .

З рис. 5 видно, що середнє значення параметра  $\alpha$  істотно зростає із збільшенням кількості елементів вихідної послідовності. Водночас коливання значень параметра  $\alpha$  із збільшенням дисперсії вихідної послідовності та зміною її середнього значення не виходить за межі статистичної похибки.

Залежність стандартного відхилення параметра  $\alpha$  від кількості елементів вихідної послідовності є більш складною (рис. 6). В області її відносно малих значень зростання кількості елементів послідовності приводить до зменшення стандартного відхилення, що є цілком природним. Але, якщо кількість елементів перевищує 1000, подальше її зростання приводить до збільшення величини  $\sigma_\alpha$ , що може бути наслідком відхилень параметра  $\alpha$  від нормального закону розподілу, які спостерігаються для тих самих значень  $n$ . Аналіз отриманих результатів свідчить, що це зумовлено появою пустих інтервалів, що пов'язано зі зростанням імовірності появи надто малих та надто високих значень елементів генерованих вихідних послідовностей. Коливання значень параметра  $\sigma_\alpha$  із збільшенням дисперсії вихідної послідовності та



зміною її середнього значення, як і для попереднього випадку, не виходить за межі статистичної похибки.

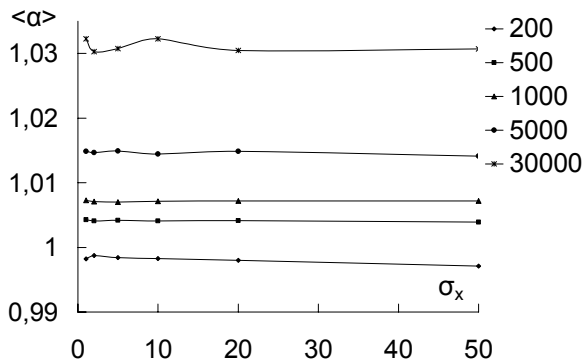


Рис. 5. Залежність середнього значення параметра  $\alpha$  від кількості елементів і стандартного відхилення вихідних послідовностей

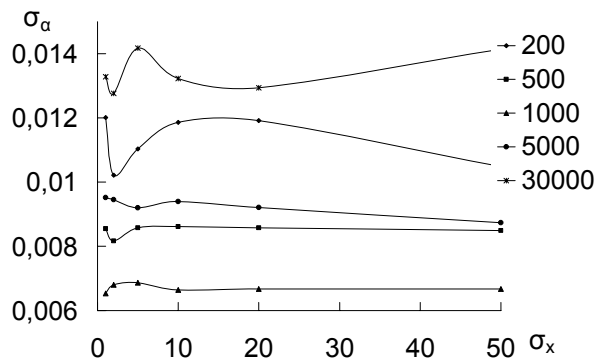


Рис. 6. Залежність стандартного відхилення параметра  $\alpha$  від кількості елементів і стандартного відхилення вихідних послідовностей

Мінімальні значення параметра  $\alpha$  закономірно підвищуються із збільшенням кількості елементів вихідної послідовності і практично не залежать від її середнього значення й стандартного відхилення, як це видно з рис. 7. Для максимального значення параметра  $\alpha$  залежність від кількості елементів послідовності є складнішою (рис. 8).

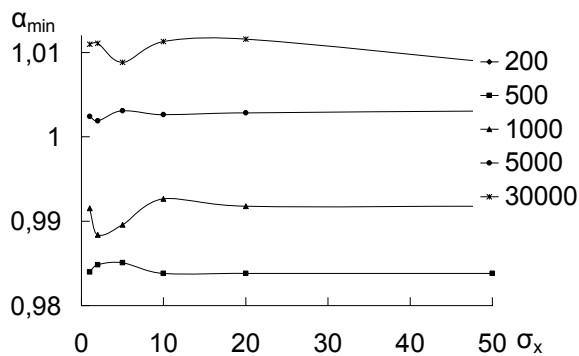


Рис. 7. Залежність мінімального значення параметра  $\alpha$  від кількості елементів і стандартного відхилення вихідних послідовностей

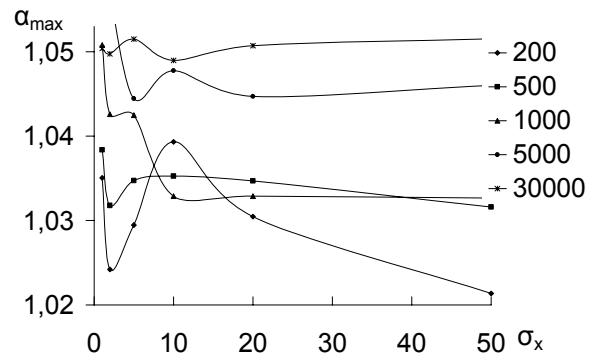


Рис. 8. Залежність максимального значення параметра  $\alpha$  від кількості елементів і стандартного відхилення вихідних послідовностей

Як підсумкову характеристику доцільно використати ймовірність того, що оцінка дисперсії за груповим методом за формулою (2) виявиться меншою, ніж оцінка за формулою (1). Для цього визначали середні значення середнього значення і стандартного відхилення параметра  $\alpha$  за всіма послідовностями, що мали однакову кількість елементів. Потім у припущенні, що значення параметра  $\alpha$  підпорядковуються нормальному закону розподілу, розраховували ймовірність того, що його значення буде меншим від одиниці. Результати наведено в табл. 1.

З таблиці видно, що із зростанням кількості елементів послідовності ймовірність того, що за груповим методом буде отримано занижену оцінку дисперсії, істотно згасає, але для послідовностей обсягом до 1000 елементів така ймовірність є досить високою. Це зумовлено як нижчими середніми значеннями параметра  $\alpha$ , так і підвищеними значеннями його середнього відхилення.



Таблиця 1

Імовірність  $p$  отримання заниженої оцінки дисперсії за формулою (2)

$n$	200	500	1000	5000	30000
$p, \%$	56,6198	31,39848	14,26861	5,683856	1,022229

### Висновки

Отримані результати дають можливість говорити, що твердження про те, що формула нахождення дисперсії завжди дає завищену оцінку дисперсії, є помилковим. Середнє значення відношення дисперсій, що розраховуються за формулами (1) та (2) збільшується із зростанням кількості елементів вибірки, за якою розраховують дисперсію і практично не залежить від середнього значення і стандартних відхилень вихідних вибірок. Імовірність того, що це значення буде менше від одиниці, зростає із зменшенням обсягу вибірки і є достатньо високою для вибірок обсягом менше ніж 1000 елементів. Така поведінка пов'язана як зі зменшенням середнього значення цього відношення при зменшенні обсягу вихідної вибірки, так і з тим, що для вибірок малого обсягу зменшення кількості елементів приводить до зростання дисперсії аналізованого параметра.

### Література

1. Бахрушин В.Є. Аналіз даних. – Запоріжжя: ГУ "ЗІДМУ", 2006. – 170 с.
2. Гайдышев И. Анализ и обработка данных: Специальный справочник. – С.Пб.: Питер, 2001. – 752 с.
3. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. – М.: Наука, 1973. – 832 с.
4. Лемешко Б.Ю. Группирование наблюдений как способ получения робастных оценок // Надежность и контроль качества. – 1997. – № 5. – С. 26–35.
5. Лемешко Б.Ю., Ванюкевич О.Н. Проверка гипотез о дисперсии при нарушении предположений о нормальности // Сб. научных трудов НГТУ. – 2002. – № 3(29). – С. 27–32.
6. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1991. – 304 с.

**Хакен Г. Информация и самоорганизация. Макроскопический поход к сложным системам. – М.: УРСС, 2005. – 248 с.**

В книге выдающегося немецкого физика-теоретика, одного из основоположников синергетики Германа Хакена развит оригинальный подход к описанию сложных макроскопических систем. Всесторонне и новаторски исследована взаимосвязь информации и самоорганизации на основе принципа максимума информационной энтропии в применении к широкому кругу неравновесных процессов. На качественно новом уровне рассмотрен синергетический подход к проблеме распознавания образов, а также сформулированы принципы синергетического компьютера. Автор изящно применяет предлагаемые им методы для исследования самых разных систем, от биологических до квантовых. Книга насыщена примерами из физики, химии, биологии, экономики, психологии. В новое издание добавлены разделы, связанные с моделированием и предсказанием процессов, основанных на неполных или зашумленных данных, а также описаны связи с теорией хаоса.