РОЗДІЛ 2
ІНФОРМАЦІЙНІ СИСТЕМИ ТА ТЕХНОЛОГІЇ

УДК 31 (075.8)

Джунь Й. В., д.ф.-м.н., професор (Міжнародний економіко-гуманітарний університет, м. Рівне).

МАТЕМАТИЧНЕ ЕСЕ ПРО ПОНЯТТЯ ОДНОРІДНИХ ВЕЛИЧИН У СТАТИСТИЦІ

Анотація. *В статті досліджено поняття «однорідна сукупність», яке часто зустрічається в підручниках з предмету «Статистика». Розкрито нечіткість його трактовок і їх відокремленість від теорії оцінок, незважаючи на те, що саме проблема оцінювання є пріоритетною при визначенні середніх величин. Піддано критиці сумнівний, невідомого походження критерій, згідно якого однорідними є спостереження у яких коефіцієнт варіації $\vartheta < 33\%$. Запропоновано математично обґрунтоване строге поняття «однорідна сукупність» – це сукупність статистичних даних, вагова функція розподілу яких є сталою. Показано, що такою властивістю володіє лише закон Гауса, отже однорідними є лише такі статистичні дані, які мають нормальний розподіл.*
Ключові слова: *однорідна сукупність, вагова функція, нормальний розподіл.*

Аннотация. *В статье исследовано понятие «однородная совокупность», которое часто встречается в учебниках по предмету «Статистика». Раскрыты нечеткость его трактовок и их отстраненность от теории оценок, хотя именно проблема оценивания является приоритетной в определении средних величин. Поддано критике сомнительный, неизвестного происхождения критерий, в соответствии с которым однородными есть наблюдения, у которых коэффициент вариации $\vartheta < 33\%$. Обосновано и предложено математически строгое понятие «однородная совокупность» – это совокупность статистических данных, весовая функция которых есть константой. Показано, что такое свойство присуще только закону Гаусса, таким образом, однородными есть только те статистические данные, которые имеют нормальное распределение.*
Ключевые слова: *однородная совокупность, весовая функция, нормальное распределение.*

Annotation. *The article examines the concept of «homogenous population» which is often found in the textbooks in «Statistics». Illegibility of its interpretations and standoffishness from the estimation theory although just the*

problem of estimation is priority in determination of half values is shown. The questionable criterion of unknown origin according to which the observations are homogenous and they have the variation coefficient is criticized. The author of the article substantiated and suggested the mathematically strict concept «homogenous population» – it is totality of statistical data and their weight function which is constant. It is shown that this property belongs only to the Gauss' law, thus, only statistical data which have normal distribution are homogenous.
Keywords: *homogeneous population, weight function, normal distribution.*

У науковій і навчально-методичній літературі зі статистики часто використовуються поняття, які мають досить неоднозначний, розпливчатий зміст. Наприклад, існує більше двадцяти різних визначень поняття «ринок праці», де кожен з авторів трактує його по-своєму, часто міняючи одну категорію на іншу, ще більш нечітку. Таку ж саму долю має трактування поняття однорідних величин у статистиці, відносно якого деякі автори відкрито пишуть про його неоднозначність і відносність [1]. Більше того, у статистичній літературі (і навіть у підручниках), «гуляють» досить сумнівні з математичної точки зору правила, наприклад, критерій однорідності. Згідно цього, так би мовити, «критерію», цілком абсурдного з точки зору теорії перевірки гіпотез, однорідними слід вважати спостереження, у яких квадратичний коефіцієнт варіації менше 33 %, хоча цілком очевидно, що однорідність даних не може визначатись ні їх дисперсією і, тим більше, значенням їх середнього арифметичного. Одним з найкращих методів, за допомогою якого можна надати будь яким категоріям чіткий і цілком однозначній зміст, є їх математизація.

Актуальність нашого дослідження полягає в розробці чіткого математичного обґрунтування поняття «однорідність даних», оскільки воно є одним із базових у статистиці. Для такого обґрунтування нами були використані основні поняття теорії оцінювання і математична теорія вагових функцій, яка детально викладена в роботі [2, с. 55].

Аналіз робіт з проблеми однорідності даних засвідчує таке. **У** статистичній навчальній літературі і в інтернет-джерелах досить часто зустрічається поняття «однорідні групи», «однорідні величини», «однорідність». Наголошується, що неоднорідні за складом сукупності потрібно розбивати на однорідні групи. Проте пояснень відносно того, що таке однорідна група, однорідна сукупність, а також рекомендацій відносно того, яким же чином можна протестувати сукупність «на однорідність», як правило, не дається. Це здається дуже дивним, оскільки поняття «якісно однорідних сукупностей» пронизує всю статистику наскрізь. Така ситуація не дозволяє якимось чином зрозуміти суть «однорідності», а й відповідно і саму суть статистичного аналізу. Якщо ж у підручнику з високими міністерськими грифами і подається якесь визначення поняття однорідності, то воно

скоріше нагадує за своїм смислом або ж щасливі догадки Аліси в країні чудес, або, в кращому випадку, нагадує мудру словесну еквілібристику Цицерона про ту саму однорідність. Наведемо кілька таких визначень: У навчальному посібнику [3, с. 89] (із серії «Вища освіта XXI століття) дано таке визначення: «Однорідні сукупності – це сукупності, елементи яких мають спільні властивості та належать до одного класу, типу». Що може дати таке досить таки туманне і нечітке визначення студенту. Можна з впевненістю сказати, що він не зрозуміє, що це таке. Візьмемо для прикладу сукупність банків України: всі вони мають спільні властивості, належать до одного класу установ і одного типу. То чи є сукупність банків, згідно цього визначення, однорідною чи не однорідною цілком не зрозуміло, тим більше, наприклад, для студента. В підручнику [4, розділ 2.1.3], зазначено, що багато дослідників мають неоднакові погляди на проблему однорідності і наводиться таке її визначення: «Поняття однорідності сукупності спостережень охоплює якісну і кількісну однорідність. Під першою треба розуміти однорідність, яка визначається однотипністю економічних об'єктів, їх однаковою якістю та певним призначенням, а під другою – однорідність групи одиниць сукупності, що визначається на основі кількісних ознак». Але як саме визначається однорідність на основі кількісних ознак, не сказано. Лише констатується: «При цьому обидва поняття діалектично взаємопов'язані і кількісна однорідність можлива лише за наявності одноякісності явищ та процесів, що «утворюють сукупність спостережень». Як бачимо, знову поняття однорідності визначається на основі ще більш нечітких категорій. На освітньому порталі «Українська педагогіка» [5] дається таке поняття: «Однорідна сукупність – якщо одна, чи декілька ознак, що вивчаються, є загальними для всіх одиниць». На сайті [6] таке: «Якісно однорідна сукупність означає, що всі її одиниці належать до одного типу явищ». Джерело [7] повідомляє: «Однорідність індивідуальних значень ознаки – це проявлення їх загальних властивостей, обумовлених основними умовами і закономірностями масового процесу, який породжує дану сукупність. У [8] зазначено: «Сукупності бувають: однорідними, якщо одна чи кілька істотних ознак, які вивчаються, є загальними для всіх одиниць, і неоднорідні – це сукупність, в яку входять явища різного типу». Досить цікавою і оригінальною є інформація про однорідність даних, яка розміщена на блозі «Статистический анализ данных в MS Excel»: «Значимость однородности в статистике трудно переоценить, так как она напрямую влияет на точность рассчитываемых показателей и качество аналитических выводов. Чем однороднее данные, тем надежнее и адекватнее реалиям результаты статистического анализа. Однородность – понятие относительное и растяжимое. Она не имеет точных границ и критериев. Под однородными данными следует понимать некоторый уровень их рассеяния, при котором рассчитываемые статистические показатели (средняя и проч.) будут давать

надежную и качественную характеристику анализируемой совокупности. Граница, отделяющая однородные данные от неоднородных, плавная и размытая. Основным мерилом разброса (и однородности) являются показатели вариации». Ось так. Майже в усіх підручниках зі статистики наводиться наступне правило, згідно якого спостереження вважаються однорідними, якщо коефіцієнт варіації сукупності:

$$\vartheta = \frac{\sigma}{\bar{x}} \cdot 100\% < 33\% \qquad (1)$$

де $\bar{x}$ – середнє вибіркове, а σ – оцінка дисперсії вибірки.

Правило (1) наводиться скрізь, як кінцева істина, без будь-якого посилання на джерело його походження чи, принаймні, на праці, де воно обґрунтовано. Те, що правило (1) безглузде з точки зору математики і теорії оцінок, – це нікого не турбує. Дослідники, які займаються конкретною науковою роботою, бачать всю сумнівність цього «правила». Наприклад, коли зважувати на аналітичних вагах зразок вагою $\sigma < 3\sigma$ то спостереження «однорідні», а коли $\sigma \geq 3\sigma$, спостереження раптом стають «неоднорідними», навіть на порчених вагах. Головне, щоб вага була велика, – хіба це не безглуздо? Однорідність даних забезпечується метрологічним налаштуванням приладу і правильною методикою вимірювань, а не вагою зразка, яка до однорідності не має ніякого відношення.

Найбільш грамотним (на жаль лише інтуїтивно) є означення однорідності сукупності, дане в роботі [9, с. 8] «Однорідність: належність всіх елементів ряду і їх вибіркових статистичних параметрів (середнього, дисперії) до однієї сукупності». Між тим існує цілком однозначне і обґрунтоване поняття однорідності даних, яке випливає з чистих і прозорих ідей засновників статистики В. Петті, А. Кетле, а найперш – сера Р. Фішера. Поняття однорідності нерозривно пов'язано з поняттям середньої величини, яку А. Кетле вважав основним методом статистичного аналізу, а А. Баулі навіть всю статистику назвав «наукою середніх величин». Згідно з теорією середніх величин А. Кетле, будь-яке явище чи процес формується під впливом двох видів факторів – постійних і випадкових. Випадкові фактори спричиняють розподіл результатів навкруги середньої. Головна мета статистики середніх – це отримати найбільш надійні значення цих середніх і їх точності. Згідно з вченням К. Пірсона [10] судить про те, які є спостереження однорідні чи ні визначаються саме особливими властивостями розподілу результатів статистичних спостережень y_i .

Метою і завданням нашого дослідження є створення однозначного і чіткого математичного обґрунтування поняття однорідних величин у статистиці, одного із найважливіших в теорії групувань.

Одним із головних завдань теорії групувань є розбиття неоднорідної за складом сукупності на однорідні групи, оскільки лише в цьому разі середні арифметичні по таким групам забезпечують ефективне оцінювання тих чи інших показників. Окрім того, щоб показати математичну суть поняття однорідності даних у статистиці, необхідно скористатись теорією вагової функції, яка розроблена з метою забезпечення ефективного оцінювання середньої величини по статистичній сукупності за допомогою наступної формули, яка передбачає застосування методу послідовних наближень [2]:

$$\bar{x} = \sum_{i=1}^{n} y_i \cdot p(x_i) / \sum_{i=1}^{n} p(x_i), \tag{2}$$

де $x_i = y_i - \bar{x}$; y_i – результати спостережень, а вагова функція:

$$p(x_i) = y_i'[y_i x_i]^{-1} = x_i^{-1} \ln' y_i . \tag{3}$$

Замінюючи у рівнянні (3) $\ln' y_i$ канонічною формою сімейств розподілів Пірсона [11, с. 101], отримуємо таку загальну аналітичну форму вагової функції (4):

$$p(x_i) = (k_0 + k_1 x + k_2 x^2)^{-1} + k_1[x(k_0 + k_1 x + k_2 x]^{-1}, \tag{4}$$

де k_0, k_1, k_2 – сталі для кожного розподілу коефіцієнти, які обчислюють за відомими формулами [11, с.101] і які залежать від асиметрії і ексцесу розподілу похибок спостережень.

Формула (4) задає нескінчену множину вагових функцій *p(x),* і кожен статистичний розподіл має свою, властиву лише йому вагову функцію. Та слід звернути особливу увагу на наступну обставину: якщо в (4) *x* = 0, то при $k_1 \neq 0$ (асиметрія розподілу), ваги *p(x)* стають нескінченними.

Таким чином, якщо статистичний розподіл результатів спостережень має значиму асиметрію, то у відповідності з теорією оцінок, середню взагалі не можна обраховувати внаслідок того, що вагова функція розподілу має вироджений характер. Іншими словами, при *A = 0* функція *p(x)* є сингулярною і будь-яке оцінювання з такою функцією є недопустимим.

Для закону Гауса (*A = 0;* $\varepsilon = \beta_2 - 3 = 0, k_1 = k_2 = 0$) вагова функція (5) набуває вигляду константи:

$$p(x_i) = 1/\sigma^2, \tag{5}$$

де σ^2 - оцінка дисперсії статистичного розподілу вибірки.

З наведеного можна зробити висновок, що статистичні дані мають однакову точність (є якісно однорідними) лише в тому разі, коли ці дані підкоряються нормальному закону розподілу. І лише такі дані можна з повним правом осереднювати, оскільки вони всі мають одинакові ваги, тобто, статистичні величини чи дані, розподіл яких не є нормальним, однорідними вважатись не можуть, так як в цьому разі, згідно з формулою (3), вони мають різну якість по точності, тому просте осереднення таких даних є недопустимим. Зазначимо, що вагова функція *p(x)*, яка підтверджує такий висновок, має розмірність оберненої дисперсії, а в загальному вигляді *p(x)* – це обернена дисперсія спостереження, похибка якого *x*.

Коротко підсумуємо основні *результати дослідження:*

1. Виходячи з формули (5), можна зробити наступний *важливий висновок: унікальною особливістю спостережень, які підкоряються закону Гауса, є те, що всі вони у цьому випадку мають однакову вагу, тобто всі спостереження, які підкоряються нормальному закону, мають однакову точність. А це означає, що лише в цьому випадку середня проста є ефективною оцінкою центра вибірки.*

2. Формула (4) показує, що за наявності асиметрії статистичних даних їх вагова функція стає виродженою. Це означає, що для таких даних саме поняття середньої величині втрачає свій смисл.

3. «Правило», згідно з яким спостереження вважають однорідними, якщо квадратичний коефіцієнт варіації менше ніж 33 %, є абсурдним, оскільки однорідність даних, як бачимо із формул (4) і (5), визначається виключно формою їх закону розподілу, а саме, його нормальністю і не залежить від того, яким є середнє значення вибірки чи її дисперсія.

Викликає здивування той факт, що це «правило» стільки років використовувалось вченими мужами вже після того, як появилась сучасна теорія оцінок і теорія вагових функцій. Скоріше всього, тут ми маємо справу з типовим академічнім обскурантизмом, який в наш час зустрічається не лише серед університетських вчених, а навіть і серед нобелівських лауреатів [2]. Може здатись, що вимоги, які ми висуваємо до поняття однорідності даних, є занадто строгими. На це можна відповісти словами А. Пуанкаре: «Математична строгість не визначає всього, але там, де її нема, – нема нічого».

За підсумками дослідження, можна дати математично точне і цілком однозначне визначення: *якісно однорідною можна вважати лише таку статистичну сукупність, яка має нормальній розподіл з досліджуваного параметру. Отже, єдиним критерієм однорідності даних є форма закону розподілу статистичної сукупності – якщо ця форма Гаусова, то лише такі дані можна вважати однорідними і ніякі інші.*

Тестування на нормальність розподілу вибірки застосовують відомим чином: або за допомогою χ^2 – квадрат критерію Пірсона, або за

допомогою перевірки статистичних гіпотез: $A = 0$, $\varepsilon = 0$. *Будь-яке значиме (по асиметрії та ексцесу) відхилення розподілу статистичної сукупності від закону Гауса означатиме різні ваги спостережень (їх різну якість і їх різну їх точність). Такі спостереження є вже якісно неоднорідними, оскільки прості середні для оцінки типового рівня ознаки в даному разі обчислювати не можна. У цьому випадку ефективними оцінками є зважені середні, які обчислюються за формулою (4).* При цьому, оцінку ваг *p(x)* ми не рекомендуємо обчислювати за формулою (4), у якій оцінки отримані методом моментів, а за формулою (3), у якій оцінки параметрів функції щільності *y* обчислюються методом максимальної правдоподібності на основі універсального закону похибок Пірсона- Джеффріса [2, с. 59].

При перевірці гіпотези на нормальність за допомогою статистичних кумулянт асиметрії і ексцесу, можна звернутись до роботи [2, с. 82], у якій наведено детальний и розширений аналіз конкретних випадків діагностики на однорідність.

1. Статистический анализ данных в MS Excel [Электронный ресурс] // Режим доступа : publications.hse.ru/books/82235802 **2.** Джунь И. В. Неклассическая теория погрешностей измерений / И. В. Джунь. Издательский дом ЕСТЕРО : Ровно, 2015. – 168 с. **3.** Уманець Т. В. Загальна теорія статистики / Т. В. Уманець. Навч. посіб. – К. : Знання, 2006. – 239 с. **4.** Поняття однорідності спостережень. – Економетрія. Навч. посіб. [Електронний ресурс] // Режим доступу : lection. com. ua/econometry/ecnavpos/ ponyattya-odnoridnostti-sposterezhen-econometriya-navchalniy-posibnik **5.** Основні поняття статистики. Ймовірність. Сукупність. Вибірка. [Електронний ресурс] // Режим доступу : ukped. com./statti/ onpd/3056-osnovni-poniattia-statystyky-ymovirnist-sukupnist- vybirka.html **6.** Характеристика та аналіз статистичних даних [Електронний ресурс] // Режим доступу : intranet. tdmu. edu/ua/data/ kafedra/ internal/ socmedic/classes_stud/uk/med/eik/pnt/ Біостатистика/4/03. **7.** Однородной совокупности, kofficienta.ru, экономические и … [Электронный ресурс] // Режим доступа : www. koefficienta.ru/enduratexter-material17modered-1188-index.html **8.** Статистика для экономистов. – Dmitry Anisimov [Электронный ресурс] // Режим доступа : anisimovdmitry. com/Documents/ MathStat Econom/mathstateconom.pdf **9.** Методические рекомендации по оценке однородности гидрологических характеристик и определению их расчетных значений по неоднородным данным. – С-Пб. : Нестор-история. 2010. – 26 с. **10.** Pearson K. On the Mathematical Theory of Errors of Judgment with special Reference to the personal Equation. / K. Pearson. // Philosophical Transaction of the Royal Sosiety of London. Ser. A, 1902, vol. 198, p. 253-296. **11.** Большев Л. Н. Таблицы математической статистики. / Л. Н. Большев, Н. В. Смирнов. – М. : Наука, 1983. – 416 с.

Рецензент: д.т.н., професор Власюк А. П.