

Джунь Й. В. д.ф.-м.н., професор (Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'ячука, м. Рівне)

НЕОКЛАСИЧНА ТЕОРІЯ ПОМИЛОК І ЇЇ ЗНАЧЕННЯ ДЛЯ СТВОРЕННЯ НОВОГО ПОКОЛІННЯ ПРОГРАМНИХ ПРОДУКТІВ В МОДУЛІ «DATA ANALYSIS»

Анотація. У статті розглянуто умови виникнення, призначення і фундаментальні принципи неокласичної теорії помилок (НТП). Показано, що НТП є необхідною еволюцією класичної теорії. Запропоновано доповнити класичний метод найменших квадратів додатковим програмним модулем для реалізації процедур НТП.

Ключові слова: теорія помилок, аналіз даних, розподіл похибок.

Аннотация. В статье рассмотрены условия возникновения, назначение и фундаментальные принципы неоклассической теории погрешностей (НТП). Показано, что НТП представляет собой необходимую эволюцию классической теории. Предложено дополнить классический метод наименьших квадратов дополнительным программным модулем для реализации процедур НТО.

Ключевые слова: теория погрешностей, анализ данных, распределение погрешностей.

Annotation. In the article the terms of rise, usage and fundamental principles of the new-classical theory of errors are considered. It is shown that non-classical theory of errors is necessary evolution of the classical theory. The author proposes to add to the classical method of the smallest squares with the supplementary software module for the non-classical theory of errors procedure realization.

Key words: theory of errors, data analysis, discrete error.

Фундаментальні принципи класичної теорії помилок (КТП) є основою сучасних процедур математичного моделювання, виявляють свою специфіку теоретичну і практичну неспроможність при значних обсягах інформації. Кембріджський професор Г. Джеффріс вперше показав, що це відбувається вже при числі багаторазових спостережень $n > 500$. У зв'язку з сказаним виникає проблема розробки нових принципів теорії помилок, адекватних практиці спостережень, які мають великі обсяги.

Мета і завдання розробленої нами неокласичної теорії помилок (НТП) якраз і полягає в розробці і обґрунтуванні цих принципів.

НТП закладає **основи** математичної обробки наукових експериментів, які мають важливе або особливо важливе значення: спостережень за рухом полюсів Землі, при лазерних локаціях ШСЗ, визначення галілеєвого прискорення, швидкості світла, сталої Планка, заряду і маси електрона, сталої надтонкої структури, тощо. Усі ці експерименти проводяться постійно і мають максимально можливу точність, величезні обсяги спостережень і характеризуються похибками з суттєвим додатнім ексцесом.

Зараз в ядерній фізиці реєструють $10^6 - 10^8$ подій за один експеримент, астрономічні каталоги охоплюють до 10^8 зірок, а класичні спостереження за рухом полюса Землі складають $1,6 \times 10^6$ широт. При такій кількості спостережень звичні нам закони математичної статистики, наприклад, закон Гауса, закон великих чисел Бернуллі, виявляють свою неспроможність. Нашим завданням було створення нового математичного апарату для аналізу надвеликих обсягів інформації. НТП якраз і присвячена розробці і обґрунтуванню нових фундаментальних принципів аналізу даних, які дозволяють подолати неспроможність постулатів КТП при великих обсягах вибірок. Розробка НТП є основним результатом докторської дисертації автора [1]. Цей новий напрямок в галузі «Data Analysis» названо по пропозиції завідуючого кафедрою вищої математики Санкт-Петербурзького інституту точної механіки і оптики професора В.Г Дегтярьова НТП.

Відомо, що КТП і метод найменших квадратів (МНК) відкриті Лежандром (1806 р.), Гаусом (1809 р.), Лапласом (1812 р.). Створений на основі КТП МНК до цього часу є одним із основних інструментів математичного моделювання і обов'язком елементом програмних продуктів з аналізу даних.

Відомо, що класичний МНК ґрунтується на наступних принципах:

- похибки спостережень підкоряються закону Гауса, якій має ексцес $E=0$;
- в результатах спостережень відсутні систематичні похибки.

В основу НТП закладено нові фундаментальні принципи:

– похибки спостережень підкоряються розподілу Пірсона VII типу з діагональною інформаційною матрицею (закону похибок Джеффіса, який вперше отримав цей розподіл);

– систему негаусових спостережень має характеризувати їх вагова функція, адаптована до розподілу похибок;

– вагова функція похибок спостережень має бути несингулярною.

Впровадження останнього принципу обумовлено тим, що в класичному МНК взагалі не враховується дія систематичних похибок. Ось що пише Гаус з цього приводу: «...потрібно з особливою силою підкреслити, що в наступних дослідженнях ми будемо говорити лише про випадкові похибки, що не мають сталої частини» [2]. Проте, кожен дослідник знає, що систематичні похибки, нехай навіть малі, є, були і завжди будуть в результатах вимірювань. Третій постулат НТП дозволяє контролювати їх вплив: дією систематичних помилок в результатах спостережень можна

знехтувати лише тоді, коли вагова функція похибок має всі ознаки несингулярності.

Перевірка спроможності закону похибок, запропонованого Джеффрісом в [3], (перший постулат НТП) здійснювалась по ініціативі і під керівництвом академіка НАНУ Е.П. Федорова. Ідея Г. Джеффріса (або сера Гарольда, як наполегливо рекомендувала називати його леді Джеффріс в листуванні з автором) полягала в наступному. Сер Гарольд запропонував для опису спостережень великих обсягів ($n > 500$) розподіл Пірсона VII типу, який є узагальненням найбільш відомих в прикладній математиці розподілів: Гауса і Стюдента. Цей розподіл має ексцеси E в межах $0 < E < \infty$ і є більш універсальним ніж закон Гауса.

На рис. подано результати експерименту, виконаного в [1], з метою перевірки ідеї Джеффріса про відповідність розподілу похибок великих обсягів кривій Пірсона VII типу. В основу перевірки покладено графік для ідентифікації типів розподілів Пірсона, взятий з [4]. Кожний розподіл похибок має на рис. три координати: ексцес E , квадрат асиметрії A^2 , обсяг n . Закону Гауса на рис. відповідає точка N , яка і є початком координат.

З рис. 1 видно, що розподіл похибок зовсім не групується в околі точки N , яка відповідає закону Гауса, а в продовж лінії кривої Пірсона VII з розмахом негаусових ексцесів, які досягають величезних значень, аж до $+\infty$.

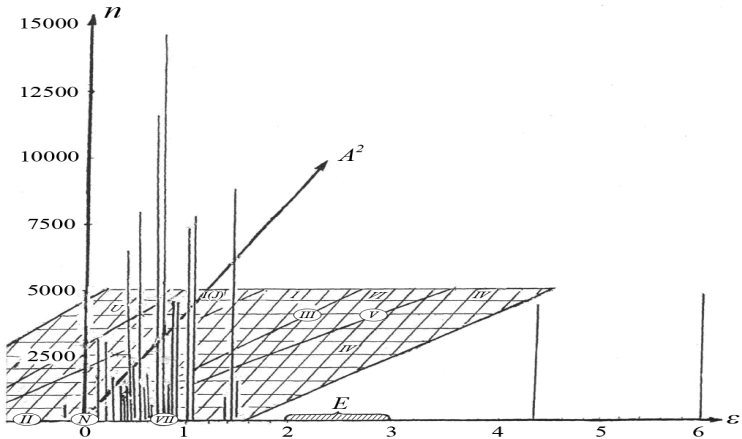


Рис. 1. Розміщення емпіричних розподілів похибок астрономічних, космічних, гравіметричних, геодезичних, економічних рядів (область E_k) на осі ексцесу ε (на графіку лінія VII, праворуч від точки N , відповідає сімейству Пірсона VII типу)

Другий з наведених вище постулатів НТП концентрує увагу дослідників тому, що додатній ексцес емпіричних розподілів похибок не є фактом, яким можна знехтувати. Він свідчить про ту надзвичайну важливу обставину, що результати спостережень мають не однакою, як при законі Гауса, а істотно різну вагу, що відображається ваговою функцією. Головна причина відхилень від закону Гауса – важкі «хвости» розподілів похибок. В реальних умовах, похибки більші «трьох сигм» виникають не з імовірністю 0,145%, як це випливає із закону Гауса, а з імовірністю приблизно в 100 раз більшою. Про це свідчать результати численних досліджень розглянутих в [5]. Цей беззаперечний факт істотно обмежує ефективність класичних процедур.

У цілому можна зробити висновок, що НТП виявилось корисним інструментом не лише при математичній обробці даних найвищої точності, але і при масових технічних, медичних, геофізичних та інших вимірах, скрізь, де виявляються негаусові похибки, а це буває не так вже і рідко. Одне з найбільш обширних досліджень дійсних розподілів похибок виконано Лабораторією прикладної математики Тартуського державного університету. Було проаналізовано 2500(!) вибірок реальних статистичних даних, – в 92% випадків гіпотеза нормальності була відхилена [6].

НТП пройшла ретельну перевірку протягом більше ніж 20 років на основі матеріалів астрометричних [1], космічних [7], гравіметричних [8], геофізичних [9], геодезичних [10], економічних [11] та інших спостережень.

Виникнення НТП обумовлено дією математичного парадоксу Ельясберга-Хампеля, згідно якого будь-яка неперервна гіпотеза про емпіричний розподіл, якою б досконалою вона не була, буде неминуче відкинута з ростом інформації.

Чи знає пересічний фахівець з інформатики чи прикладної математики, який використовує програмні продукти для «Data Analysis» про цей парадокс, про те, що в середньому в 90-92%, а то і в 98% випадків, фундаментальні принципи МНК не відповідають дійсності. Можна дати відповідь – ні. Ці фахівці будують зараз свої моделі навіть не здогадуючись про низьку технологічність своїх методів, іноді не розуміючи того, що моделювання відбувається в сингулярних областях. Щоб цього не було, необхідно створити програмні продукти нового покоління, в яких були б задіяні принципи НТП і логіку яких розумів би кожен програміст, а споживачі були впевнені в тому, що обробка даних виконана на сучасному рівні.

Не слід думати, що НТП є якоюсь революційною теорією, яка відкидає КТП і МНК. Ніякі революції, згідно з теорією систем, нічого не дають. НТП є еволюційним продовженням КТП. Її процедури не змінюють звичних схем класичного моделювання. При застосуванні методів НТП класичний МНК є необхідним першим етапом моделювання, а вже потім, при необхідності, настає черга рафінованих процедур НТП.

У чому є небезпека, яка загрожує новітнім і перспективним розробкам? В [12] є повідомлення, що по аналізу даних надруковано 10^6 статей і книг, з яких лише 10% актуальні. Тобто, лівова частка публікацій є звичайним непотребом. Технології НТП друкуються в дорогих або зарубіжних виданнях, і недоступні нашим студентам. З іншого боку, які б важливі наукові результати не були б отримані, якщо вони не будуть відомі студентам, то ці результати можуть сезнути в болоті наукової макулатури. Тому потрібно найскоріше віддрукувати НТП і ввести її як доповнення до курсів КТП. НТП вже написана, та ненадрукована. Її видання в Україні оцінюється в 75.000 грн. Для цього коштів нема. Є, правда, пропозиції від LAP Lambert Academic Publishing (Німеччина) безкоштовно надрукувати НТП на російській мові. Але європейське видання НТП буде коштувати 120-150 євро і навряд буде доступне нашим студентам. Та є надія, що НТО викличе інтерес в Україні (і навіть серед посадових осіб) і буде надрукована у нас. В усякому разі розробка НТО є одним із найбільших вагомих наукових досягнень університету і присвячується його 20- річчю.

1. Джунь И. В. Математическая обработка астрономической и космической информации при негауссовых ошибках наблюдений. Автореферат дис. докт. физ – мат. наук. – К.: ГАО АНУ. 2. Гаусс К. Ф. Избранные геодезические сочинения. Том 1. Способ наименьших квадратов. Под ред. Г. В. Багратуни. Пер. лат. и нем. – М.: Изд. геод. лит., 1957. – 234 с. 3. Jeffreys H. Theory of Probability. Sec. Edition. Oxford. 1940. 4. Большев Л. Н. Таблицы математической статистики / Л. Н. Большев, А. В. Смирнов. – М.: ВЦ АН СССР, 1968. – 474 с. 5. Ершов А. А. Стабильные методы оценки параметров (обзор) / А. А. Ершов // Автоматика и телемеханика, 1978, № 8 (66) – 100 с. 6. Орлов А. И. Часто ли распределение результатов наблюдений является нормальным ? / А. И. Орлов // Заводская лаборатория. – 1991, № 7. – С. 64–66. 7. Dzhun I. V. Pearson's Distribution of Type VII of the Errors of Satellite Laser Ranging Data // Kinematics and Physics of Celestial Bodies. – Allerton Press. Inc. / New York, 1991, vol.7, p.p. 74-84. 8. Джунь И. В. Особенность закона распределения результатов баллистических измерений ускорения силы тяжести / И. В. Джунь, Г. П. Арнаутов, Ю. Ф. Стусь // Изд. МГК при Президиуме АН СССР и НПО “Нефтегеофизика” Повторные гравиметрические наблюдения. М.: 1984, с.87-100. 9. Джунь. И. В. О некоторых фундаментальных вопросах математической обработки геофизической информации / И. В. Джунь // Геодинамическое исследование в Украине: сб. науч. трудов НАНУ. Ин. геофизики им. С.И. Субботина. Наукова думка, 1995 – 210с. 10. Джунь. И. В. Об аппроксимации плотности вероятности некоторых рядов ошибок геодезических измерений распределением Пирсона VII типа / И. В. Джунь // Геодезия и аерофотосъемка. 1989. – № 6. – С. 43–48. 11. Dzhun I.V., Gazda V. About Distribution of Stock Index Returns Fluktuations // Business Reviev. Scientific Journal of Faculty of Bussines Ekonomics of the Universiti of Economics in Bratislava with a seat in Kosice, 2002, vol. I. № 2, p. 20-27. 12. Орлов А.И. Высокие статистические технологии / А. И. Орлов // Заводская лаборатория, 2003. – Т. 69. – № II. – С.55-60.

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