DOI 10.32782/2786-8559/2023-3-17
УДК 330.4:519.85

Юськів Богдан Миколайович
доктор політичних наук, професор,
професор кафедри економіки та управління бізнесом,
Рівненський державний гуманітарний університет
ORCID: https://orcid.org/0000-0001-7621-5954

Пляшко Ольга Степанівна
кандидат економічних наук, доцент,
доцент кафедри економіки та управління бізнесом,
Рівненський державний гуманітарний університет
ORCID: https://orcid.org/0000-0001-7202-1036

Хомич Сергій Васильович
кандидат економічних наук, доцент,
доцент кафедри економіки та управління бізнесом,
Рівненський державний гуманітарний університет
ORCID: https://orcid.org/0000-0002-8737-1992

ЗАРОДЖЕННЯ «АНАЛІЗУ ДАНИХ» ЯК НАУКОВОЇ ДИСЦИПЛІНИ

Стаття висвітлює еволюцію аналізу даних від традиційної статистики до науки про дані. Починаючи з твердження Пітера Хьюбера про емпіричний характер аналізу даних, де дослідник наголошує, що цей етап розвитку не можна визначити як нову наукову парадигму, але як певну тенденцію, яка об'єднується під назвою «наука про дані». Основний акцент робиться на внеску Джона Т'юкі, який першим висловив ідеї, що лягли в основу аналізу даних. Робота розкриває концепції «підтверджувального» та «експлораторного» аналізу даних, визначає їхні цілі та різницю, а також відзначає важливість чергування цих етапів у процесі дослідження. Принципи Т'юкі для сучасного аналізу даних, такі як «максимальне проникнення в дані» та «візуалізація закономірностей», розглядаються як ключові підходи для виявлення нових знань. Роботи Т'юкі викликали значні дебати серед статистиків, а його погляди на аналіз даних шокували академічне співтовариство. Докладно розглядається вплив робіт Т'юкі на розвиток науки про дані протягом півстоліття, включаючи коментарі відомого статистика Пітера Губера. Заклики Т'юкі до реформування статистики та його погляди на важливість ставлення правильних запитань і отримання приблизних відповідей наголошують на його важливості в контексті аналізу даних і науки про дані. Важливий акцент робиться на впливі обчислювальних середовищ на розвиток аналізу даних. Зазначається, що реальний прогрес в розумінні поняття «аналіз даних» був стимульований кодом і обчислювальними середовищами. Вказується на роль різних статистичних пакетів та програмних середовищ, таких як BMDP, SPSS, SAS, Minitab, S, STATA і R у розвитку аналізу даних. Вивчається їхній вплив за допомогою аналізу частоти слів у літературі. Зазначається, що сьогодні R є домінуючим середовищем програмування в академічній статистиці з великою кількістю прихильників. Завдяки роботі зі скриптами можна точно кодифікувати кроки обчислень. Ці зміни викликали зміну у правилах гри, і тепер вираз «науковий підхід до аналізу даних» став більш очевидним, відповідаючи твердженню Дж. Т'юкі щодо можливостей вивчення аналізу даних як науки.

Ключові слова: *аналіз даних, наука про дані, Data Mining, R, SPSS, статистика.*

Bohdan Yuskiv, Olha Pliashko, Sergii Khomych
Rivne State University of the Humanities

THE EMERGENCE OF "DATA ANALYSIS" AS A SCIENTIFIC DISCIPLINE

The article describes the evolution of data analysis from traditional statistics to data science. Starting with Peter Huber's assertion about the empirical nature of data analysis, where the researcher emphasizes that this stage of development cannot be defined as a new scientific paradigm but rather as a tendency unified under the name "data science". The main focus is on the contributions of John Tukey, who first expressed ideas that laid the foundation for data analysis. The article explores the concepts of "confirmatory" and "exploratory" data analysis, defines their goals and differences, and emphasizes the importance of

alternating between these stages in the research process. Tukey's principles for contemporary data analysis, such as "maximum insight into the data" and "visualization of patterns", are considered key approaches for discovering new knowledge. Tukey's works sparked significant debates among statisticians, and his views on data analysis shocked the academic community. The impact of Tukey's works on the development of data science over half a century is examined, including comments from the renowned statistician P. Huber. An essential emphasis is placed on the influence of computational environments on the development of data analysis. The role of various statistical packages and software environments, such as BMDP, SPSS, SAS, Minitab, S, STATA, and R, in the evolution of data analysis is discussed. Their impact is assessed through the analysis of word frequencies in the literature, highlighting that R is currently the dominant programming environment in academic statistics with a large number of enthusiasts. The use of scripts to precisely codify computation steps is noted, and these changes are seen as altering the rules of the game, making the expression "scientific approach to data analysis" more evident, aligning with Tukey's assertion about the possibilities of studying data analysis as a science.

Keywords: *data analysis, data science, data mining, R, SPSS, statistics.*

Вступ. За останній рік однією зі світових топновин є розвиток штучного інтелекту та суміжних з ним технологій. Перш за все, мова йде про так звані мовні моделі та програмні продукти, що будовані на їх основі, зокрема ChatGPT. Часто моделі та методи, пов'язані зі штучним інтелектом, машинним навчанням, комп'ютерним баченням сприймають як щось абсолютно нове, що з'явилося порівняно недавно. Частково це дійсно так, адже ці наукові напрямки дуже інтенсивно розвиваються і як наслідок відбуваються справжні «прориви». Проте в їх основі лежать підходи, що були закладені статистикою, теорією ймовірності та іншими «класичними» напрямками науки. Метою дослідження є історичний погляд на становлення та трансформацію аналізу даних як наукової дисципліни, її взаємозв'язок зі статистикою, розуміння чинників, що впливали на процес трансформації термінів та підходів, пов'язаних з аналізом даних.

Матеріали та методи. При написанні статті використовувалися як емпіричні, так і теоретичні методи дослідження. Емпіричний підхід передбачав збір та аналіз конкретних фактів і даних, в той час як теоретичний метод базувався на використанні теоретичних концепцій та моделей для розуміння явища. Крім того, застосовувалися методи статистичного аналізу для виявлення закономірностей та трендів у досліджуваних даних, що надавало можливість отримати узагальнені дані з відповідною перевіркою.

Результати. Як констатує відомий швейцарський статистик Пітер Хьюбер (Peter J. Huber) [1], еволюційний шлях від аналізу даних до науки про дані розпочався в середовищі статистиків і математиків у 1962 році.

Одним з перших і найголовніших висновків полягав у тому, що «аналіз даних за своєю суттю є емпіричною наукою» [2, с. 63]. Хоча сам по собі аналіз даних не можна розглядати як нову чітко визначену наукову парадигму. Він являє собою швидше нову назву, котра об'єднує в собі низку тенденцій, що з'явилися і проявилися в обробці даних, починаючи від другої половини XX століття. Однак, як виявилося згодом, за цими тенденціями стояла спільна філософія, котрій невдовзі дали назву – наука про дані (Data Science).

Першим, хто виразно заговорив про аналіз даних, був відомий фахівець у сфері математичної статистики Джон Т'юкі (John Tukey). Свої ідеї щодо нових тенденцій він виклав у 1962 році у статті під назвою «Майбутнє аналізу даних» [3].

Т'юкі Дж. говорив про те, що точність і строгість математичних основ статистики не допомагають розв'язувати реальні життєві проблеми і що треба дати даним говорити самим за себе. Тому дослідник запровадив термін «аналіз даних» і обґрунтував необхідність розрізнення двох видів аналізу, що однаковою мірою важливі (рівнозначні) для отримання нових знань на основі дослідження емпіричного матеріалу. Перший різновид, у основі якого покладена традиційна/класична статистична теорія перевірки гіпотез, дослідник назвав «підтверджувальним аналізом даних» («confirmatory data analysis»). Доповненням до нього є «експлораторний (або розвідувальний/пошуковий) аналіз даних» («exploratory data analysis»). Цей підхід не просто доповнює «підтверджувальний аналіз даних», а в сенсі цілей є його протилежністю. «Експлораторний аналіз даних» має за мету не перевірку гіпотез, а їх формулювання, точніше: висування гіпотез про причини досліджуваного явища, припущень, на яких ґрунтуватиметься подальший статистичний висновок, і відтак – забезпечення бази для подальшого збирання даних у процесі досліджень.

Т'юкі Дж. також запропонував принципи, які стали підґрунтям того, що називають сучасний аналіз даних:

– максимальне «проникнення» в дані;
– виявлення основних структур даних;
– вибір найважливіших змінних;
– виявлення відхилень і аномалій;
– перевірка основних гіпотез (припущень);
– розроблення початкових моделей;
– візуалізація закономірностей, прихованих у даних.

Дослідник особливо наголошував на важливості чергування експлораторного і підтвер-

джувального етапів аналізу в процесі виявлення нового знання: спочатку (перший етап) експлораторний (пошуковий) аналіз використовується для «генерації ідей», а потім ці ідеї на інших вибірках/множинах даних перевіряються і приймаються для роботи щонайбільше як гіпотеза.

Вказана стаття не пройшла осторонь товариства статистиків, представником якого був сам Т'юкі Дж. Як зауважує Девід Донохо (David Donoho), «стаття Джона була опублікована в 1962 році в «Анналах математичної статистики», центральному місці для математично передових статистичних досліджень того часу. Інші статті, що в той час з'являлися в цьому журналі, були математично точними і містили визначення, теореми та доведення. Натомість стаття Джона Т'юкі була свого роду публічною сповіддю, яка пояснювала, чому він вважав такі дослідження надто вузькоспрямованими, можливо, марними або шкідливими, а сферу досліджень статистики необхідно значно розширити і переорієнтувати» [4, с. 749]. Стаття викликала не просто дебати, а й бурхливі суперечки серед представників класичної математичної статистики.

Т'юкі Дж. глибоко шокував своїх читачів (академічних статистиків) уже своїми вступними фразами: «Довгий час я вважав себе статистиком, який цікавиться висновками від часткового до загального. Але коли я спостерігав за розвитком математичної статистики, у мене з'явилися причини дивуватися і сумніватися… Врешті-решт, я прийшов до висновку, що мій головний інтерес полягає в аналізі даних, який включає в себе, серед іншого, процедури аналізу даних, методи інтерпретації результатів таких процедур, способи планування збору даних, щоб зробити їх аналіз простішим, точнішим або більш точним, а також весь апарат і результати (математичної) статистики, які застосовуються для аналізу даних...» [2, с. 2].

У роботі науковець обґрунтовано наголошує: «Найважливіша максима, якої слід дотримуватися при аналізі даних, і якої, здається, багато статистиків уникають, полягає в наступному: набагато краще мати приблизну відповідь на правильне запитання, яке часто є розпливчастим/невизначеним, ніж точну відповідь на неправильне запитання, яке завжди можна уточнити» [2, с. 13–14]. Тобто, на думку Т'юкі Дж., надто велика увага до статистичної перевірки гіпотез, властива класичному статистичному аналізу тільки збіднює статистичну методологію і може навіть давати хибні результати через неправильну постановку запитань. Тим самим науковець фактично закликав реформувати академічної статистики.

Майже через 50 років відомий статистик Пітер Губер (Peter Huber) прокоментував наслідки цієї статті [3, с. 749]: «Т'юкі у своїй надзвичайно впливовій статті переосмислив наш предмет... [Стаття] ввела термін «аналіз даних» як назву того, чим займаються прикладні статистики, відмежувавши цей термін від формального статистичного висновку. Але насправді, як зізнався Т'юкі, він «розтягнув термін за межі його філології» до такої міри, що він охопив усю статистику». Таким чином, бачення Т'юкі включило статистику в більш широке поняття. Основним твердженням Т'юкі було те, що ця нова сутність, яку він назвав «аналіз даних», є новою наукою, а не просто галуззю математики.

Т'юкі Дж. також визначив чотири рушійні сили (чинники розвитку) в новій науці:

1) формальні теорії статистики;

2) прискорення розвитку комп'ютерів і пристроїв відображення;

3) поява у багатьох галузях все більших і більших масивів даних;

4) акцент на кількісній оцінці в дедалі ширшому спектрі дисциплін.

Як пише Донохо Д. [3, с. 749], список Джона 1962 року напрочуд сучасний і охоплює всі чинники, які сьогодні згадуються в прес-релізах, що рекламують сучасні ініціативи в галузі науки про дані. Хоча шокуючим на той час був п. 1, який означав, що статистична теорія є лише (незначною!) частиною нової науки. Цю нову науку порівнювали з усталеними науками і ще більше обмежували роль статистики в ній: аналіз даних – це дуже складна сфера. Він повинен адаптуватися до того, що люди можуть і повинні робити з даними. У тому сенсі, що біологія складніша за фізику, а поведінкові науки ще складніші за них, тому цілком ймовірно, що загальні проблеми аналізу даних є більш складними, ніж у всіх трьох згадуваних науках. Аналіз даних може багато чого отримати від формальної статистики, але теоретична статистика може відігравати лише часткову роль у його розвитку. Зауважимо, що цю думку Т'юкі Дж. виразно повторив через 6 років, давши таку назву своїй новій книзі – «Аналіз даних, який включає статистику» [4].

Заклики Т'юкі Дж. щодо зміни підходу до статистики не були сприйняті одразу – пройшло десять-двадцять років, поки науковці і практики не тільки реально оцінили і прийняли бачення свого видатного колеги, а й стали розвивати його ідеї. Серед тих, хто зробив вагомий внесок у цей розвиток, Донохо Д. називає відомих статистиків Джона Чемберса (John Chambers), Джеффі Ву (Jeff Wu), Вільяма Клівленда (William S. Cleveland), які незалежно один від одного знову закликали академічну статистику розширити свої межі за межі класичної сфери теоретичної статистики. Хоча ці заклики мали відносно невеликий видимий ефект до 2000 року.

Чемберс Дж., один з розробників мови S для статистики та аналізу даних у Bell Labs, у 1993 році опублікував есе під провокаційною

назвою «Більша чи менша статистика, вибір для майбутніх досліджень» [5]. Науковець закликав приділяти більше уваги підготовці та представленню даних, а не статистичному моделюванню. Такі дослідження зосереджуватимуться на можливостях, що надаються новими типами даних і новими типами їхнього представлення. Чемберс Дж. чітко заявив, що розширена сфера буде більшою навіть за аналіз даних. Це глибше і ширше, ніж бачення Т'юкі у 1962 році.

Ву Дж. після своєї інавгурації на посаду професора статистики в Мічиганському університеті прочитав інавгураційну лекцію під назвою «Статистика – наука про дані?». Він запропонував перейменувати статистику на науку про дані, а статистиків – на науковців, що займаються наукою про дані. Передбачаючи сучасні магістерські курси з науки про дані, він навіть згадав ідею нового магістерського ступеня, в якому близько половини курсів були б поза межами статистики. Він охарактеризував статистичну роботу як трилогію зі збору даних, моделювання та аналізу даних і прийняття рішень. Фактично для нової передбачуваної галузі саме Ву Дж. запропонував привабливу назву – «наука про дані» (Data Science).

Клівленд В. розробив багато цінних статистичних методів і способів відображення даних, працюючи в Bell Labs. Його стаття 2001 року під назвою «Наука про дані: План дій для розширення технічних напрямів у галузі статистики» [6] була присвячена академічним відділам статистики і пропонувала план переорієнтації їхньої роботи. В анотації статті записано: «План дій для розширення технічних галузей статистики зосереджується на аналітиці даних. Він визначає шість технічних напрямів роботи для університетської кафедри і пропонує конкретний розподіл ресурсів, призначених для досліджень у кожній галузі та для курсів у кожній галузі. Цінність технічної роботи оцінюється тим, наскільки вона приносить користь аналітику даних, прямо чи опосередковано. Цей план також може бути застосований до державних дослідницьких лабораторій і корпоративних дослідницьких організацій». У вступі до своєї статті Клівленд В. констатує, що «...[результати в] науці про дані слід оцінювати за тим, якою мірою вони дозволяють аналітику вчитися на даних... Інструменти, які використовує аналітик даних, приносять пряму користь. Теорії, які слугують основою для розробки інструментів, дають непряму користь». Дослідник запропонував шість напрямків дій, визначивши частку кожного :

– міждисциплінарні дослідження (25%);
– моделі та методи для даних (20%);
– обчислення з даними (15%);
– педагогіка (15%);
– оцінювання інструментів (5%);
– теорія (20%).

Сьогодні ідеї Т'юкі Дж. здобули загальне визнання [7], його науковий доробок розглядають як підґрунтя становлення аналізу даних із його логічним продовженням – інтелектуального аналізу даних [8; 9]. Як зауважує українська соціологиня Кислова О. [10], дискусії щодо науковості «неточного» експлораторного підходу завершено, а ортодоксальні погляди на роль математики в статистичному аналізі замінилися визнанням аналізу даних як повноправної наукової дисципліни, що має: 1) певні особливості, які відрізняють її від прикладної математичної статистики, 2) потенціал подальшого розвитку – перетворення на інтелектуальний аналіз даних (завдяки комп'ютеризації наявних методів).

Визначальним чинником цього визнання стали винаходи та розроблення обчислювальних середовищ для аналізу даних. Як пише Д. Донохо [3, с. 750], «зробити діяльність під назвою «аналіз даних» більш конкретною і видимою, зрештою, стимулював код, а не слова».

До таких середовищ належать ранні статистичні пакети BMDP, SPSS, SAS і Minitab, які беруть свій початок з обчислень на мейнфреймах кінця 1960-х років, а також пакети S, ISP, STATA і R, які беруть свій початок з епохи міні-комп'ютерів/персональних комп'ютерів. Щоб кількісно оцінити важливість цих пакетів використано сервіс Google's N-grams viewer і побудовано графік частоти слів SPSS, SAS, Minitab у книгах англійською мовою з 1960 по 2000 рік на основі сервісу Google Books; а для порівняння маємо графік частот біграм «аналіз даних» і «статистичний аналіз». Виявляється (рис. 1), що SAS і SPSS є більш поширеними термінами в англійській мові в цей період, ніж «аналіз даних» чи «статистичний аналіз», зокрема майже вдвічі частіше, ніж «аналіз даних» [3, с. 750].

Також ми порівняли популярність термінів «Data Science», «Data Mining», «Data Analysis», «Statistical Analysis» в англомовних книгах сервісу Google Books за 1960–2019 рр. Цікаво, що термін «Data Analysis» залишається досить популярним, «Statistical Analysis» поступово виходить з літературного обігу, поряд з тим зростає популярність «Data Science» (рис. 2).

Чемберс Дж. та його колега Рік Беккер (Rick Becker) з Bell Labs у середині 1970-х роках розробили середовище кількісних обчислень «S». Це середовище пропонувало мову для опису обчислень, а також низку базових статистичних засобів та інструментів візуалізації. У 1990-х роках ця система дістала своє продовження як система R – проект з відкритим вихідним кодом, який швидко завоював популярність. Сьогодні R є домінуючим середовищем кількісного програмування, що використовується в академічній статистиці і має вражаючу кількість прихильників в Інтернеті.

Середовища кількісного програмування, подібні до R, запускають «скрипти», які точно коди-

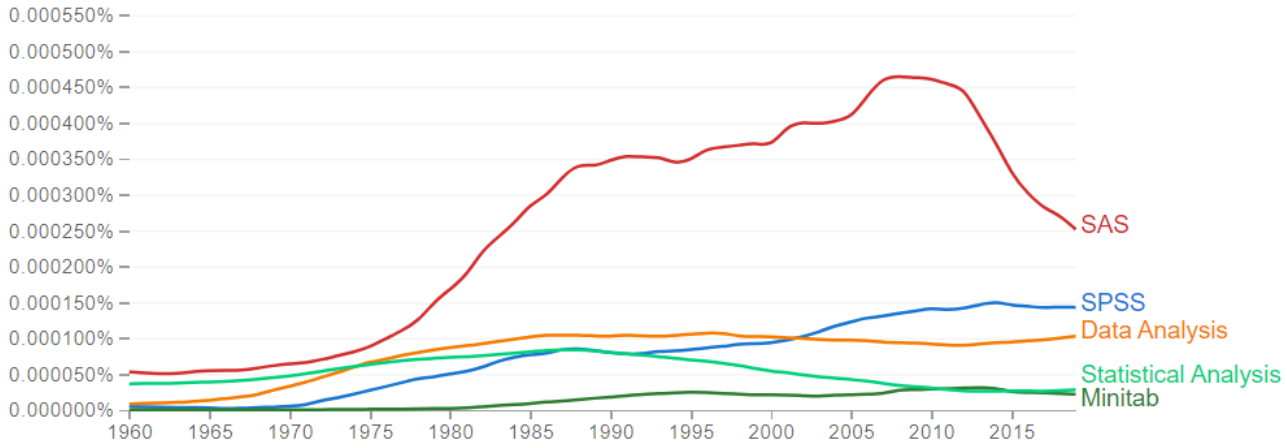

Рисунок 1 – Графік частот слів SPSS, SAS, Minitab, пар слів «Data Analysis» і «Statistical Analysis»

Джерело: [11]

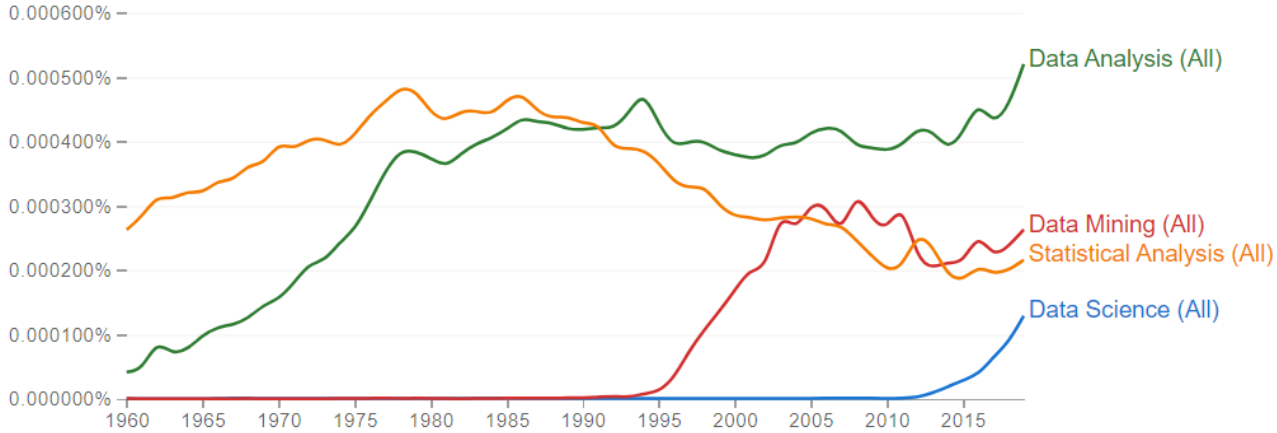

Рисунок 2 – Графік частот пар слів «Data Analysis», «Statistical Analysis», «Data Science» та «Data Mining»

Джерело: [12]

фікують кроки обчислень. Такі скрипти сьогодні часто називають робочими процесами. Це змінило правила гри. Тепер є сенс говорити про науковий підхід до поліпшення аналізу даних, а твердження Дж. Т'юкі про те, що вивчення аналізу даних може бути наукою, стало самоочевидним [3, с. 750–751].

Висновки. За результатами проведеного дослідження можна зробити висновок, що аналіз даних як наукова дисципліна пройшов тривалий шлях становлення від статистики та суміжних з нею наукових напрямків. Важливий внесок у становлення аналізу даних як окремої наукової парадигми роблять програмні технології та середовище програмування, в першу чергу мова R. Поряд з аналізом даних розвиваються й інші поняття, такі як наука про дані (Data Science), видобуток даних (Data Mining).

Література:

1. Huber P.J. Data Analysis: What Can Be Learned From the Past 50 Years. John Wiley & Sons, 2011.
2. Tukey J.W. The future of data analysis. *Annals of Mathematical Statistics*. 1962. Vol. 33. № 1. P. 1–67.
3. Donoho D. 50 Years of Data Science. *Journal of Computational and Graphic Statistics*. 2017. No 26(4). P. 745–766. DOI: https://doi.org/10.1080/10618600.2017.1384734 (дата звернення: 08.11.2023).
4. Mosteller F., Tukey J.W. Data Analysis, Including Statistics. Handbook of Social Psychology / Eds. G. Lindzey, E. Aronson. Vol. 2. Reading, MA : Addison-Wesley, 1968. P. 80–203.

5. Chambers J.M. Greater or Lesser Statistics: A Choice for Future Research. *Statistics and Computing*. 1993. No. 3. P. 182–184.
6. Cleveland W.S. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review*. 2001. No. 69. P. 21–26.
7. Brillinger D.R., Fernholz L.T., Morgenthaler S. The Practice of Data Analysis: Essays in Honor of John W. Tukey. Princeton, New Jersey : Princeton University Press, 1997. 352 p.
8. Dempster A.P. John W. Tukey as «philosopher». *Annals of Mathematical Statistics.* 2002. Vol. 30. № 6. P. 1619–1628. URL: http://surl.li/ntixf (дата звернення: 08.11.2023).
9. Kafadar K. John Tukey and Robustness. *Statistical Science*. 2003. Vol. 18. № 3. P. 319–331. URL: http://surl.li/ntixn (дата звернення: 08.11.2023).
10. Кислова О.Н. Интеллектуальный анализ данных: история становления термина. *Український соціологічний журнал*. 2011. № 1–2. С. 83–94. URL: http://surl.li/ntixs (дата звернення: 08.11.2023).
11. Google's N-grams viewer. URL: http://surl.li/ntiyc (дата звернення: 08.11.2023).
12. Google's N-grams viewer. URL: http://surl.li/ntiyj (дата звернення: 08.11.2023).

References:

1. Huber P. J. (2011) *Data Analysis: What Can Be Learned From the Past 50 Years.* John Wiley & Sons.
2. Tukey J. W. (1962) The future of data analysis. *Annals of Mathematical Statistics*, vol. 33, no. 1, pp. 1–67.
3. Donoho D. (2017) 50 Years of Data Science. *Journal of Computational and Graphic Statistics,* no. 26(4), pp. 745–766. DOI: https://doi.org/10.1080/10618600.2017.1384734
4. Mosteller F. & Tukey J. W. (1968) *Data Analysis, Including Statistics*. Handbook of Social Psychology / Eds. G. Lindzey, E. Aronson. Reading, MA: Addison-Wesley, vol. 2.
5. Chambers J. M. (1993) Greater or Lesser Statistics: A Choice for Future Research. *Statistics and Computing,* no. 3, pp. 182–184.
6. Cleveland W. S. (2001) Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review*, no. 69, pp. 21–26.
7. Brillinger D. R., Fernholz L. T. & Morgenthaler S. (1997) *The Practice of Data Analysis: Essays in Honor of John W. Tukey*. Princeton, New Jersey: Princeton University Press.
8. Dempster A. P. & John W. (2002) Tukey as "philosopher". *Annals of Mathematical Statistics,* vol. 30, no. 6, pp. 1619–1628. Available at: http://surl.li/ntixf
9. Kafadar K. (2003) John Tukey and Robustness. *Statistical Science*, vol. 18, no. 3, pp. 319–331. Available at: http://surl.li/ntixn
10. Kyslova O. (2011) Intelektualnyy analiz danykh: istoriya rozvytku termina [Data mining: the history of the term]. *Ukrayinskyy sotsiolohichnyy zhurnal,* no. 1-2, pp. 83–94. Available at: http://surl.li/ntixs
11. Google's N-grams viewer. Available at: http://surl.li/ntiyc
12. Google's N-grams viewer. Available at: http://surl.li/ntiyj

Стаття надійшла до редакції 21.11.2023 р.