медицина, науковий потенціал та історія науки, рівень досліджень в нашій країні наближається до світового. Проте дослідженню економічних дисциплін в нашій країні приділяється невиправдано мала увага.

1. *Presenting* INIS. — Vienna: International Atomic Energy Agency,1999.
2. www.iaea.org.

Є.П. Удалов
(Київський національний університет імені Тараса Шевченка)
Ю.А. Хомич,
наук.співроб. (Центр досліджень науково-технічного потенціалу та історії науки ім.Г.М. Доброва НАН України)

Аналіз вибіркових даних при оцінюванні наукового потенціалу і характер статистичних властивостей вербальних моделей

Історично доведено, що класичною областю прикладної математичної статистики є математичні методи вибіркових досліджень. Найбільше застосування вони знаходять у медицині й соціології. Починаючи з 1970—1975 років у нашій країні розвиток сучасних вибіркових методів, зокрема статистики об'єктів нечислової природи, стимулювався запитами соціологічних і експертних досліджень [1]. О.І.Орловим та його співавторами розроблено нові підходи, сформульовано постановки, запропоновано алгоритми аналізу різнотипних даних (які включають значення кількісних і якісних ознак), отримано теореми про властивості цих алгоритмів, спроможність оцінок і т.п., причому зведення теоретичних результатів було видано у вигляді монографії [2].

Незважаючи на успіхи теоретичних досліджень у статистиці, до методик проведення конкретних вибіркових досліджень і аналізу отриманих даних за допомогою відповідних програмних продуктів справа не дійшла. Перехід до ринкової економіки в Україні та на теренах колишнього СРСР, що супроводжується різким спадом виробництва, високим рівнем інфляції, дефіцитом державного бюджету, зменшення попиту на дослідження і розробки з боку промисловості негативним чином позначилися на стані науки. Негативні процеси, що відбуваються у вітчизняній науці, вимагають нових підходів як до методів збору даних про її стан (фінансування, матеріально-технічна база, результативність, кадри, інфраструктура і т.д.), так і до методів аналізу отриманих даних.

Віддзеркаленням процесів, які мають місце в останні роки в науці, можуть стати кілька цифр, що відображають фінансування НАН України: 2001 р. — 413,4; 2002 р. — 449,3; 2003 р. — 588,6; 2005 р. — 914,9 млн. грн. (18,12 млн. $) [3]. Як бачимо, фінансування зростає, але, враховуючи коєфіцієнти інфляції, зростання не таке вже і велике. Збереженню потенціалу української науки приділяється увага міжнародної наукової громадськості, окремих країн, міжнародних організацій, зокрема Українського науково-технологічного центру. Загальне фінансування його проектів у 2004 р. склало 17,9 млн. $. Зазначимо, що провідні світові фірми вкладають у

наукові розробки значно більші кошти (в 2005 р. фірма «Simens» — 6500 млн. $, «Samsung» — 4500 млн. $, «Microsoft» — 5800 млн. $ [4]).

Коли обговорюються різноманітні задачі вивчення науки і керування нею, дуже важливими є вихідні методологічні принципи, однакове розуміння й оцінка досліджуваних процесів. Показовою є дискусія, в якій один з авторів як основний показник використовував продуктивність праці науковця [5], а інший — фондоємність наукової продукції [6], що явно ускладнювало взаєморозуміння. Суттєво, що в обох статтях широко використовувалися як статистичні, так і експертні дані. Безперечним видається висновок, що статистичні дані про науковий потенціал — база для теоретичного і прикладного наукознавства.

1. Підходи до статистики вербальних об'єктів, можливість їхньої алгоритмізації і застосування для аналізу статистичних даних

З початку 70-х років XX ст. набуває активного розвитку статистика об'єктів нечислової природи (ОНП), відома також як статистика нечислових даних. У розвитку цього порівняно нового напрямку прикладної математичної статистики пріоритет належить російським ученим.

На сьогодні статистика ОНП в теоретичному плані досить добре розвинена, основні ідеї, методи і підходи описано і вивчено в математичному аспекті, доведено досить багато теорем. Однак теорія поки що недостатньо апробована практично.

Висловлюється припущення, що при аналізі даних про науковий потенціал методи статистики ОНП виявляться найбільш корисними, оскільки істотна частина даних має нечисловий, якісний (вербальний) характер.

Головним елементом математичної статистики є вибірка. Ймовірнісна теорія статистики показує, що вибірка — це сукупність незалежних однаково розподілених випадкових елементів. Розглянемо сутність цих елементів. Класична математична статистика подає елементи вибірки як числа, багатовимірний статистичний аналіз — як вектори. А в нечисловій статистиці елементами вибірки є ОНП, які не піддаються математичним діям — діленню на числа чи складанню. Інакше кажучи, вербальні об'єкти лежать у просторах, що не мають векторної структури.

ОНП — об'єкти, які недоцільно описувати числами, зокрема елементами нелінійних просторів. Прикладами є бінарні співвідношення, такі як розбивки, ранжування, толерантності та ін., результати парних і множинних порівнянь, вимір у шкалах, відмінних від абсолютних, множини, нечіткі множини. Наведемо приклади ОНП, що мають якісні ознаки: стать людини чи тип наукової організації, результат віднесення об'єкта до одної із заданих градацій (категорій); сукупність людей, що займаються визначеної працею (фізичною, розумовою та ін.); слово, пропозиція, текст, шрифт, що у пам'яті комп'ютера кодуються за допомогою цифр 0 і 1, але не стають від цього числами; розбивки сукупностей на групи об'єктів, подібних між собою (кластери); ранжування — упорядкування експертами наукових проектів за ступенями переваги і т.п. Інтервальні дані теж можна розглядати як приклад об'єктів нечислової природи.

Розглянемо принципову новизну статистики ОНП. У математичній статистиці зазвичай використовується операція додавання (віднімання) для розрахунку вибіркових характеристик розподілу, таких як вибіркове середнє арифметичне, вибіркова дисперсія й т.п. У регресійному аналізі та інших областях цієї дисципліни постійно використовуються суми. Апарат математичної статистики оперує великими числами, тому закони великих чисел, центральна гранична теорема й інші теореми націлені на вивчення сум.

У вербальній статистиці не можна використовувати операцію додавання, оскільки елементи вибірки лежать у просторах, де немає операції додавання. Тому методи обробки вербальних даних засновані на принципово іншому математичному апараті — застосуванні різних відстаней у просторах ОНП.

Оскільки нечислові дані складають близько 90% даних у соціології і 70% в економіці, теоретичні дослідження в статистиці нечислових даних дозволяють одержати нові результати у тій центральній області економетрики, в якій роботи російських вчених мають пріоритет на світовому рівні [7].

1.1. Середні дані, отримані за теоретичним і практичним планом

Із самого початку необхідно однозначно з'ясувати, яким чином проводиться визначення середніх величин для ОНП. Класична математична статистика вводить середні величини за допомогою операцій додавання (вибіркове середнє арифметичне, математичне очікування) чи упорядкування (вибіркова і теоретична медіани). У просторах довільної природи середні значення не можна визначити за допомогою операції додавання. Доводиться вводити теоретичні та емпіричні середні як рішення екстремальних задач. Теоретичне середнє (у класичному змісті) — це рішення задачі мінімізації математичного очікування відстані від випадкового елементу зі значеннями в розглянутому просторі до фіксованої крапки цього простору. Для середнього, отриманого практичними діями, тобто емпіричного середнього, математичне очікування береться за емпіричним розподілом, тобто береться сума відстаней від деякої крапки до елементів вибірки і потім мінімізується по цій крапці. При цьому і емпіричне, і теоретичне середні як рішення екстремальних задач можуть бути не єдиними елементами простору, а складатися із множин таких елементів, які можуть виявитися і порожніми. Проте О.І. Орлову вдалося сформулювати і довести закони великих чисел для середніх величин, визначених зазначеним чином, тобто збіжність емпіричних середніх до теоретичного приросту обсягу вибірки [2, 8, 9]. Ним з'ясовано, що методи доказу законів великих чисел допускають істотно більш широку область застосування, ніж та, для якої вони були розроблені. А саме, вдалося вивчити асимптотику рішень екстремальних статистичних задач, до яких, як відомо, зводиться більшість постановок прикладної статистики [9]. Зокрема, крім законів великих чисел, встановлено і множину оцінок мінімального контрасту, в тому числі оцінок максимальної правдоподібності та робастних оцінок. Подібні оцінки вивчені й в інтервальній статистиці.

1.2. Розділення об'єктів нечислової природи на види

Значний інтерес становлять результати, пов'язані з конкретними областями статистики ОПН, зокрема зі статистикою нечітких множин, випадковими множинами (слід зазначити, що теорія нечітких множин у визначеному змісті зводиться до теорії випадкових множин [8, 10]), з непараметричною теорією парних порівнянь, аксіоматичним введенням метрик у конкретних просторах ОПН.

Сучасні методи класифікації, в тому числі типології, дуже важливі для аналізу даних про наукові організації України, їх науковий потенціал. Проблемами теорії і практики класифікації в нашій країні займаються багато науковців. Але, мабуть, найбільш природно ставити і вирішувати задачі класифікації в рамках статистики об'єктів нечислової природи. Зазначене має відношення як до розпізнавання образів із вчителем (дискримінантний аналіз), так і розпізнавання образів без вчителя (кластерний аналіз). Сучасний стан дискримінантного і кластерного аналізів відбито з погляду статистики ОПН у працях [11—13].

Статистичні методи аналізу нечислових даних пристосовані для застосування в соціології і наукознавстві, оскільки в цих областях до 90% даних є нечисловими.

1.3. Основи теорії вимірів

Спочатку розглянемо перехід від соціологічного завдання до математичного, а саме до однієї з означених постановок проблеми однозначності в репрезентативній теорії виміру [14, 15]. Почнемо з розгляду конкретного соціологічного дослідження.

При вивченні привабливості різних професій для випускників шкіл [16] був складений список з 30 професій. Опитуваних просили оцінити кожну із цих професій одним із балів 1, 2, ..., 10 за правилом: чим більше подобається, тим вищий бал. Для одержання соціологічних висновків необхідно було дати єдину оцінку привабливості певної професії для сукупності випускників шкіл. Як така оцінка у праці [16] використовувалося середнє арифметичне балів, виставлених професіям опитаними школярами. Зокрема, фізика одержала середній бал 7,69, а математика — 7,50. Відповідно до логіки [16] фізика як професія краща, ніж математика.

Однак було відзначено [17], що цей висновок суперечить даним праці [18], згідно з якими ленінградські (петербурзькі) школярі середніх класів більше люблять математику, ніж фізику. Обговоримо одне з можливих пояснень цього протиріччя, що полягає в методиці обробки даних, застосованих у праці [16].

Справа, мабуть, в тому, що бали 1, 2, ..., 10 введені дослідником-соціологом суб'єктивно. Якщо одна професія оцінена в 10 балів, а друга — в 2, то із цього не можна виснувати, що перша рівно в 5 разів привабливіша другої. Інший колектив соціологів міг би прийняти іншу систему балів, наприклад 1, 4, 9, 16, ..., 100. Природно припустити, що впорядкування професій по привабливості, властивій школярам, не залежить від того, якою системою балів їм запропонує користуватися соціолог. Раз так, то розподіл професій за градаціями десятибальної системи не зміниться, якщо перейти до іншої системи балів за допомогою строго зростаючої функції $\gamma: K^1 \to K^1$. Якщо $x_1, x_2, ..., x_n$ — відповіді n випускників шкіл, що стосуються математики, а $y_1, y_2, ..., y_n$ — фізики, то після переходу до нової системи балів відповіді щодо математики матимуть вигляд $\gamma(x_1), \gamma(x_2), \ldots, \gamma(x_n)$, а щодо фізики — $\gamma(y_1), \gamma(y_2), \ldots, \gamma(y_n)$.

Нехай єдина оцінка привабливості професії обчислюється за допомогою функції $f(x_1, x_2, \ldots, x_n)$. Які ж вимоги природно накласти на функцію $f: K^n \to K^1$, щоб отримані з її допомогою висновки не залежали від системи балів, обраної соціологом?

Єдина оцінка обчислювалася для того, щоб порівнювати професії по привабливості. Тому зажадаємо стійкості результату порівняння: нерівність

$$f(x_1, x_2, \ldots, x_n) < f(y_1, y_2, \ldots, y_n) \quad (1)$$

справедлива тоді й тільки тоді, коли справедлива нерівність

$$f(\gamma(x_1), \gamma(x_2), \ldots, \gamma(x_n)) < \\ < f(\gamma(y_1), \gamma(y_2), \ldots, \gamma(y_n)), \quad (2)$$

причому однозначність нерівностей (1) і (2) є при будь-яких x_i, y_i і γ. Які f стійкі щодо порівняння? Відповідь на це питання було дано у праці [17]. Зокрема, з'ясувалось, що середнім арифметичним, як у праці [16], користуватися не можна, а членами варіаційного ряду і тільки ними — можна і необхідно [2].

1.4. Вербальні об'єкти як статистичні дані

Математична статистика має найпоширеніший об'єкт вивчення — вибірку $x_1, x_2, ..., x_n$, тобто сукупність результатів n спостережень. Різні області статистики подають результат спостереження як число, або кінцевовимірний вектор, або функцію. Відповідно проводиться розподіл математичної статистики: одно-

вимірна статистика, багатовимірний статистичний аналіз, статистика тимчасових рядів і випадкових процесів. У статистиці ОНП як результати спостережень розглядаються об'єкти нечислової природи, зокрема перерахованих вище видів — виміру в шкалах, відмінних від абсолютної, бінарні відношення, вектори з 0 і 1, множини, нечіткі множини. Вибірка може складатися з *n* ранжувань і *n* толерантностей, *n* множин, *n* нечітких множин і т.п. [12, 19].

Тому відзначимо необхідність розвитку методів статистичної обробки «різнотипних даних», обумовлену великою роллю в прикладних дослідженнях «ознак змішаної природи» [20].

Результат спостереження стану об'єкта найчастіше являє собою вектор, в якого частина координат вимірювана по шкалі найменувань, частина — по порядковій шкалі, частина — по шкалі інтервалів і т.д. Статистичні методи орієнтовані звичайно або на абсолютну шкалу, або на шкалу найменувань (аналіз таблиць спряженості), а тому найчастіше непридатні для обробки різнотипних даних. Є й більш складні моделі різнотипних даних, наприклад, коли деякі координати вектору спостережень описуються нечіткими множинами [21].

Для позначення подібних некласичних результатів спостережень в 1979 р. й запропоновано збірний термін — ОНП (об'єкти нечислової природи) [20]. Термін «нечисловий» означає, що структура простору [22], в якому лежать результати спостережень, не є структурою дійсних чисел, векторів або функцій, вона взагалі не є структурою лінійного (векторного) простору. При розрахунках об'єкти числової природи, зрозуміло, зображуються за допомогою чисел.

З метою «стандартизації математичних знарядь» доцільно розробляти методи статистичного аналізу даних, придатні одночасно для всіх перерахованих вище видів результатів спостережень. Крім того, в процесі розвитку прикладних досліджень виявляється необхідність використання нових видів ОНП, відмінних від розглянутих вище, наприклад у зв'язку з розвитком статистичних методів обробки текстової інформації [23]. *Тому доцільно ввести ще один вид ОНП — об'єкти довільної природи (ОДП), тобто елементи множини, на які не накладено ніяких умов (крім «умов регулярності», необхідних для справедливості доказуваних теорем).* Інакше кажучи, у цьому випадку передбачається, що результати спостережень, як правило, це елементи вибірки, і вони лежать у довільному просторі *X*. Для одержання теорем необхідно зажадати, щоб *X* задовольняло деяким умовам, наприклад було топологічним простором. Відомо, що ряд результатів математичної статистики отриманий саме в такій постановці. Так, при вивченні оцінок максимальної правдоподібності елементи вибірки можуть лежати в просторі довільної природи. Це не впливає на міркування, оскільки в них розглядається лише залежність щільності ймовірності від параметру. Методи класифікації, що використовують лише відстань між об'єктами, які класифікуються, можуть застосовуватися до сукупностей об'єктів довільної природи, аби тільки в просторі, де вони лежать, була задана метрика. Ціль статистики ОНП полягає в тому, щоб систематично розглядати методи статистичної обробки даних як довільної природи, так і таких, які являють собою зазначені вище конкретні види ОНП, тобто методи опису даних, оцінювання й перевірки гіпотез. Погляд із загальної точки зору дозволяє одержати нові результати й в інших областях математичної статистики.

1.5. Використання вербальних об'єктів при формуванні математичної моделі

Використання ОНП часто породжене бажанням обробляти більш об'єктивну, більше звільнену від погрішностей інформацію. Як показали численні досліди, людина більш правильно (і з меншими утрудненнями) відповідає на пи-

тання якісного, наприклад порівняльного, характеру, ніж кількісного. Так, їй легше сказати, яка із двох гир важча, ніж вказати їх вагу в грамах. Інакше кажучи, використання об'єктів нечислової природи — засіб підвищити стійкість економетричних і математичних моделей реальних явищ. Спочатку конкретні області статистики об'єктів нечислової природи (а саме прикладна теорія вимірів, нечіткі й випадкові множини) були розглянуті в монографії [2] при аналізі часткових постановок проблеми стійкості математичних моделей соціально-економічних явищ і процесів до припустимих відхилень вихідних даних і передумов моделі. Тому була зрозуміла необхідність проведення робіт із розвитку статистики об'єктів нечислової природи як самостійного наукового напрямку.

Обговорення почнемо зі шкал виміру. Як відомо, єдність мір і точність вимірів вивчає метрологія. Таким чином, репрезентативна теорія вимірів — частина метрології. Методи обробки даних повинні бути адекватні щодо припустимих перетворень шкал виміру згідно репрезентативної теорії вимірів. Однак встановлення типу шкали, тобто задання групи перетворень Ф, — справа фахівця відповідної прикладної області. Так, оцінки привабливості професій ми вважали вимірюваними в порядковій шкалі [2]. Однак окремі соціологи не погоджувалися із цим, вважаючи, що випускники шкіл користуються шкалою з більш вузькою групою припустимих перетворень, наприклад інтервальною шкалою. Очевидно, ця проблема стосується не математики, а наук про людину. Для її вирішення може бути поставлений досить трудомісткий експеримент. Поки ж він не поставлений, доцільно приймати порядкову шкалу, тому що це гарантує від можливих помилок.

Як ми вже відзначали, номінальні й порядкові шкали великою мірою поширені не тільки в соціально-економічних дослідженнях. Вони застосовуються в медицині, мінералогії, географії й т.д. Нагадаємо, що шкалою інтервалів вимірюють величину потенційної енергії або координату крапки на прямій, на якій не відзначені ні початок, ні одиниця виміру; за шкалою відносин — більшість фізичних одиниць (масу тіла, довжину, заряд), а також ціни в економіці. Час вимірюється по шкалі різниць, якщо рік приймаємо природною одиницею виміру, і по шкалі інтервалів у загальному випадку. У процесі розвитку відповідної області знання тип шкали може змінюватися. Так, спочатку температура вимірювалася по порядковій шкалі (холодніше — тепліше), потім по інтервальній (шкали Цельсія, Фаренгейта, Реомюра) і, нарешті, після відкриття абсолютного нуля температур — по шкалі відносин (шкала Кельвіна). Слід зазначити, що серед фахівців іноді є розбіжності з приводу того, за якими шкалами варто вважати вимірюваними ті або інші реальні величини.

Відзначимо, що термін «репрезентативна» використовувався, щоб відрізнити розглянутий підхід до теорії вимірів від класичної метрології, а також від робіт А.М.Колмогорова й А. Лебега, пов'язаних із виміром геометричних величин, від «алгоритмічної теорії виміру» та інших наукових напрямків.

Необхідність використання в математичних моделях реальних явищ таких об'єктів нечислової природи, як бінарні відносини, множини, нечіткі множини, коротко була показана вище. Тут же звертається увага, що аналізовані в класичній статистиці результати спостережень також «не зовсім числа». А саме будь-яка величина X вимірюється завжди з деякою погрішністю ΔX і результатом спостереження є

$$Y = X + \Delta X. \qquad (3)$$

Як ми вже відзначали, погрішностями вимірів займається метрологія. Відзначимо справедливість наступних фактів:

а) для більшості реальних вимірів неможливо повністю виключити систематичну помилку, тобто $M(\Delta X) \neq 0$;

б) розподіл ΔX у переважній більшості випадків не є нормальним [2];

в) вимірювану величину X і погрішність її виміру ΔX звичайно не можна вважати незалежними випадковими величинами;

г) розподіл погрішностей оцінюється за результатами спеціально проведених вимірів, отже, повністю відомим вважати його не можна; найчастіше дослідник володіє лише границями для систематичної погрішності й оцінками таких характеристик випадкової погрішності, як дисперсія або розмах.

Наведені факти показують обмеженість області застосовності розповсюдженої моделі погрішностей, в якій X і ΔX розглядаються як незалежні випадкові величини, причому ΔX має нормальний розподіл з нульовим математичним очікуванням.

Строго кажучи, результати спостереження завжди мають дискретний розподіл, оскільки описуються числами, в яких небагато значущих цифр (звичайно від 1 до 5). Виникає дилема: або визнати, що безперервні розподіли — внутріматематична фікція, і припинити ними користуватися, або вважати, що безперервні розподіли мають «реальні» величини X, які спостерігаються із принципово неперебороною погрішністю ΔX. Перший вихід у наш час недоцільний, бо він вимагає відмови від більшої частини розробленого математичного апарату. Із другого випливає необхідність вивчення впливу неперебороних погрішностей на статистичні висновки.

Погрішності ΔX можна враховувати або за допомогою ймовірнісної моделі (ΔX — випадкова величина, яка має функцію розподілу, що загалом залежить від X), або за допомогою нечітких множин. У другому випадку приходимо до теорії нечітких чисел і до її часткового розділу — статистики інтервальних даних.

Інше джерело появи погрішності ΔX пов'язане з прийнятою в конструкторській і технологічній документації системою допусків на контрольовані параметри виробів і деталей, з використанням шаблонів при перевірці та контролю якості продукції тощо [24]. У цих випадках характеристики ΔX визначаються не властивостями засобів виміру, а застосовуваною технологією проектування й виробництва. У термінах прикладної статистики сказаному відповідає угруповання даних, за якого ми знаємо, якому із заданих інтервалів належить спостереження, але не знаємо точного значення результату спостереження. Застосування угруповання може дати економічний ефект, оскільки найчастіше легше (в середньому) встановити, до якого інтервалу належить результат спостереження, ніж точно виміряти його.

1.6. Об'єкти нечислової природи як результат статистичної обробки даних

Об'єкти нечислової природи з'являються не тільки на «вході» статистичної процедури, але й в процесі обробки даних і на «виході» як підсумок статистичного аналізу.

Розглянемо найпростішу прикладну постановку завдання регресії [2]. Вихідні дані мають вигляд $(x_i, y_i) \in R^2$, $i = 1, 2, ..., n$. Ціль полягає в тому, щоб із достатньою точністю описати y як багаточлен (поліном) від x, тобто модель має вигляд

$$y_i = \sum_{k=0}^{m} a_k x_i^k + \varepsilon_i, \quad i = 1, 2, \ldots, n, \quad (4)$$

де m — невідомий ступінь поліному; a_0, a_1, a_2, ..., a_m — невідомі коефіцієнти багаточлену; ε_i, $i = 1, 2, ..., n$ — погрішності, які для простоти приймемо незалежними і які мають той самий нормальний розподіл.

Тут наочно проявляється одна із причин живучості статистичних моделей на основі нормального розподілу. Такі моделі, хоча й, як правило, неадекватні реальній ситуації [2], з математичної точки зору дозволяють проникнути глибше в суть досліджуваного явища. Тому вони придатні для первісного аналізу ситуації, як і в розглянутому випадку. Подальші наукові дослідження повинні бути спря-

мовані на зняття нереалістичного припущення нормальності й перехід до непараметричних моделей погрішності.

Розповсюджена процедура відновлення залежності за допомогою багаточлену така: спочатку намагаються застосувати модель (4) для лінійної функції ($m = 1$), при невдачі (неадекватності моделі) переходять до багаточлена другого порядку ($m = 2$), якщо знову невдача, то беруть модель (2) з $m = 3$ і т.д. (адекватність моделі перевіряють за *F-критерієм* Фішера).

Обговоримо властивості цієї процедури в термінах прикладної статистики. Якщо ступінь полінома заданий ($m = m_0$), то його коефіцієнти оцінюють методом найменших квадратів, властивості цих оцінок добре відомі. Однак в описаній вище реальній постановці m теж є невідомим параметром і підлягає оцінці. Таким чином, потрібно оцінити об'єкт (m, a_0, a_1, a_2, ..., a_m), множина значень якого можна описати як $R^1 \cup R^2 \cup R^3 \cup \ldots$. Це об'єкт нечислової природи, звичайні методи оцінювання для якого незастосовні, тому що m — дискретний параметр. У розглянутій постановці розроблені до теперішнього часу методи оцінювання ступеня полінома мають в основному евристичний характер. Властивості описаної вище розповсюдженої процедури розглянуті в [2], де знайдено граничний розподіл оцінок цього параметра, який виявився геометричним. Відзначимо, що для ступеня багаточлена давно запропоновані методи оцінки [25].

У більш загальному випадку лінійної регресії дані мають вигляд

$$(y_i, X_i),\ i = 1, 2, \ldots, n,$$

де $X_i = (x_{i1}, x_{i2}, \ldots, x_{iN}) \in R^N$ — вектор предикторів (факторів, що пояснюють змінні), а модель така:

$$y_i = \sum_{j \in K} a_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n, \quad (5)$$

де K — деяка підмножина множини $\{1,2,...,n\}$; ε_i — ті ж, що й у моделі (4); a_j — невідомі коефіцієнти при предикторах з номерами з K.

Модель (4) зводиться до моделі (5), якщо

$$x_{i1}=1,\ x_{i2}=x_i,\ x_{i3}=x_i^2,$$
$$x_{i4}=x_i^3, \ldots, x_{ij}=x_i^{j-1}, \ldots$$

У моделі (4) є природний порядок введення предикторів у розгляд відповідно до зростання ступеня, а в моделі (5) природного порядку немає, тому тут наявна довільна підмножина множини предикторів. Є тільки частковий порядок: чим потужність підмножини менша, тим краще. Модель (5) особливо актуальна в технічних дослідженнях. Вона застосовується в завданнях керування якістю продукції й інших техніко-економічних дослідженнях, у медицині, економіці, маркетингу й соціології, коли з великої кількості факторів, що приблизно впливають на досліджувану змінну, треба відібрати по можливості найменше число значимих факторів і з їх допомогою сконструювати прогнозуючу формулу (5).

Завдання оцінювання моделі (5) розбиваються на два послідовні завдання: оцінювання множини K — підмножини множини всіх предикторів, а потім — невідомих параметрів a_j. Методи вирішення другого завдання добре відомі й докладно вивчені (звичайно використовують метод найменших квадратів).

Набагато гірше справа з оцінюванням об'єкта нечислової природи K. Як ми вже зазначали, існують методи, в основному евристичні, які найчастіше не є обґрунтованими. Навіть саме поняття обґрунтованості в цьому випадку вимагає спеціального визначення. Нехай K_0 — дійсна підмножина предикторів, тобто підмножина, для якої справедлива модель (5), а підмножина предикторів K_n — її оцінка. Оцінка K_n називається обґрунтованою, якщо

$$\lim_{n \to 0} Card(K_n \Delta K_0) = 0, \quad (6)$$

де Δ — символ симетричної різниці множин; Card(K) означає число елементів множини K, а межа розуміється в сенсі збіжності по ймовірності.

Завдання оцінювання в моделях регресії, таким чином, розбивається на два: оцінювання структури моделі й оцінювання параметрів при заданій структурі. У мо-

делі (4) структура описується ненегативним цілим числом *m*, у моделі (5) — множиною *K*. Структура — об'єкт нечислової природи. Завдання її оцінювання складні, тоді як завдання оцінювання чисельних параметрів при заданій структурі добре вивчені, розроблені ефективні методи (у розумінні прикладної математичної статистики).

Така ж ситуація й в інших методах багатовимірного статистичного аналізу: у факторному аналізі (включаючи метод головних компонентів) і багатовимірному шкалуванні, в інших оптимізаційних постановках проблем прикладного багатовимірного статистичного аналізу [19].

Перейдемо до об'єктів нечислової природи на «виході» статистичної процедури. Приклади численні: розбивки — підсумок роботи багатьох алгоритмів класифікації, зокрема алгоритмів кластер-аналізу; ранжування — результат упорядкування професій по привабливості або результати автоматизованої обробки думок експертів — членів комісії з підведення підсумків конкурсу наукових праць. (В останньому випадку використовуються ранжування зі зв'язками; так, в одну групу, найбільш численну, попадають роботи, що не одержали нагород). Із всіх об'єктів нечислової природи, мабуть, найбільш часті на «виході» дихотомічні дані [2]: прийняти або не прийняти гіпотезу, прийняти або забракувати партію продукції тощо. Результатом статистичної обробки даних може бути множина, наприклад зона найбільшого ураження при аварії, або послідовність множин, наприклад «середньовимірний» опис поширення пожежі (розділ 4 [2]). Нечіткою множиною Е.Борель [11] ще на початку XX ст. пропонував описувати уявлення людей про кількість зерен, що утворюють «купу».

За допомогою нечітких множин формалізуються значення лінгвістичних змінних, які виступають як підсумкова оцінка якості систем автоматизованого проектування, сільськогосподарських машин, побутових газових плит, надійності програмного забезпечення або систем керування. Можна констатувати, що всі види вербальних об'єктів можуть з'являтися «на виході» статистичного дослідження.

2. Приклади можливих застосувань сучасних методів аналізу вибіркових даних у задачах вивчення наукового потенціалу і керування ним

У силу цілого ряду причин — високої динаміки змін, потреби в оперативній інформації для прийняття рішень і т.п., — виникає гостра необхідність регулярного проведення оперативних обстежень наукових організацій і соціологічних опитувань вчених. Ця задача може бути успішно вирішена тільки шляхом проведення вибіркових обстежень. Наведемо кілька прикладів можливих застосувань сучасних статистичних методів аналізу вибіркових даних у задачах вивчення і керування науковим потенціалом.

Вивчення показників наукової діяльності та експертно-статистичний підхід до побудови інтегральних показників. Статистичний і експертний аналіз показників у сфері науки, виявлення їх рейтингу: які показники є найбільш значимими з врахуванням загальноприйнятих міжнародних стандартів у статистиці науки [25]. Яким чином треба оцінювати ефективність науки? Якою мірою загальноприйняті показники результативності науки (публікації, патенти, ліцензії, індекси цитування і т.п.) відбивають реальну ефективність науки?

Кластеризація наукових організацій, виділення типів. При обговоренні питань фінансування, організаційних перетворень, перспективного розвитку необхідно до різних типів наукових організацій підходити диференційовано. Наскільки прийнята типологія наукових організацій відповідає тим чи іншим аналітичним задачам? Скільки типів існує реально? Яка природна типологія наукових організацій? Якщо така типологія надто складна, чи не вдасться її спростити, розглядаючи наукові організації з визначених практичних позицій: з погляду нау-

кового рівня, віддачі, фінансування, перспективності та виживання і т.д.

Становить також інтерес структура наукових рад, комісій та інших форм діяльності вчених, особливо в ситуації, коли через такі структури йде фінансування.

Рейтинги проектів, наукових організацій та ін. У рамках досить однорідної сукупності проектів НДР, заявок на гранти, наукових організацій і т.д. за допомогою методів багатовимірного статистичного аналізу можна виявити основні напрямки варіації.

Однак головний фактор всупереч розповсюдженому способу інтерпретації результатів факторного аналізу (чи методу головних компонентів) не завжди відповідає осі «ефективність — неефективність». Проте ідея рейтингу (інтегрального показника якості) заслуговує проробления. Можна вказати кілька підходів. Для вирішення цієї задачі можна використовувати три варіанти експертно-статистичного методу інтегральної оцінки перспективності наукового напрямку чи напрямку роботи наукової організації:

- за інтегральним показником (лінійна функція, параметри якої — показники щодо науки, а значення коефіцієнтів виходить від експертів);
- за вибірками, отриманими від кваліфікованих експертів;
- за допомогою оцінки параметрів інтегрального показника по вибіркам (власне експертно-статистичний метод).

Розрахунок і короткострокове прогнозування показників щодо науки. Найважливіше значення для прийняття управлінських рішень у сфері науки має прогнозування таких показників, як зайнятість, середня зарплата в секторі «Наука і наукове обслуговування» та ін., особливо з врахуванням ефекту зміни макроекономічних показників (зокрема індексу інфляції, курсу долара й інших валют) і впливу тих чи інших заходів (економічних, законодавчих, податкових, митних і т.д.), які починаються на державному рівні.

Застосування сучасних статистичних методів аналізу нечислових і інтервальних даних. Використання статистики об'єктів нечислової природи у вибіркових обстеженнях, зокрема регресійного аналізу у просторах різнотипних ознак (об'єктів нечислової природи), дасть можливість оцінити ефективність фінансування, а статистика інтервальних даних дозволить врахувати неминучі неточності в наявних даних.

Виділення «груп ризику». Окремий випадок обговорюваних вище задач — виділення (за формальними — звітними — ознаками) наукових організацій, саме існування яких виявляється під сумнівом у найближчому майбутньому.

Прогноз «виживаності» НДІ може бути побудований за допомогою вибірок на основі непараметричних оцінок щільності в просторі різнотипних ознак, частина координат яких — кількісні ознаки, а частина — якісні.

Існує багато інших цікавих проблем [27, 28]. Наприклад, виявлення циклічності розвитку науково-технічного потенціалу країни, вивчення динаміки реальної і формальної структури науки [29], форм примітивізації наукової діяльності та наукової продукції в умовах різкого спаду виробництва і скорочення фінансування наукової праці, проблеми відображення в суспільній свідомості й у самосвідомості наукового співтовариства специфіки наукової діяльності (включаючи аналіз розповсюджених догм) та ін.

Застосування сучасних статистичних методів у вибіркових дослідженнях наукових організацій дозволить одержувати результати, цікаві з теоретичної точки зору і корисні для практики управлінських рішень, впливаючих на розвиток науки.

1. *Орлов А.И.* Нечисловая статистика // Наука и технология в России. — 1994. — № 3 (5). — С.7—8.
2. *Анализ* нечисловой информации в социологических исследованиях / Под ред. В. Г. Андреенкова, А. И. Орлова, Ю. Н. Толстовой. — М.: Наука, 1985. — 220 с.

3. *Національна* академія наук України. Короткий річний звіт 2004 р. (http://www.nas.gov.ua/ua/mainold.html).

4. *Компания* и рынки // Дело. — 2005. — 10 ноября (№ 17). — С. 12.

5. *Орлов А.И.* Социологический прогноз развития российской науки на 1993—1995 годы // Наука и технология в России. — 1993. — № 1.

6. *Страхов В.Н.* Нужны ли подобные прогнозы? // Там же.

7. *Научно*-техническая и инновационная политика. Российская Федерация. Т. 1. Оценочный доклад / Организация экономического сотрудничества и развития, 1994. — 124 с.

8. *Орлов А.И.* Устойчивость в социально-экономических моделях. — М.: Наука,1979 — 296 с.

9. *Орлов А.И.* Асимптотика решений экстремальных статистических задач // Анализ нечисловых данных в системных исследованиях. — М.: ВНИИСИ, 1982. — С. 4—12. — (Тр. ВНИИСИ. — 1982. — Вып.10).

10. *Орлов А.И.* Задачи оптимизации и нечеткие переменные. — М.: Знание, 1980. — 64 с.

11. *Орлов А.И.* Классификация объектов нечисловой природы на основе непараметрических оценок плотности // Проблемы компьютерного анализа данных и моделирования: Сб.науч.ст. — Минск: Белорус. гос. ун-т, 1991. — С. 141—148.

12. *Орлов А.И.* Некоторые вероятностные вопросы теории классификации // Прикладная статистика. — М.: Наука, 1983. — С. 166—179.

13. *Орлов А.И.* Заметки по теории классификации // Социология: методология, методы, математические модели. — 1992. — № 2. — С.28—50.

14. *Джини К.* Средние величины. — М.: Статистика, 1970. — 556 с.

15. *Кетепу J.* // Pacif. J. Math . — 1959. — Vol. 9, № 4.— P. 1179—1189.

16. *Кендалл М.Дж., Стьюарт А.* Статистические выводы и связи. — М.: Наука, 1973. — 899 с.

17. *Карапетян К.А., Чахмахнян А.А.* // Тез. докл. Второй всесоюз. школы-семинара «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа». — М.: ЦЭМИ АН СССР, 1983. — Т. 2. — С. 10—18.

18. *Фоменко А.Г.* // Проблемы устойчивости статистических моделей: Тр. семинара. — М.: ВНИИСИ, 1984. — С. 154—177.

19. *Орлов А.И.* Организационные методы управления наукой и статистика объектов нечисловой природы // Тез. докл. Всесоюз. симпоз. «Медицинское науковедение и автоматизация информационных процессов». — М., 1984.— С. 215—216.

20. *Дылько Т.Н.* // Вестн. Белорус. гос. ун-та. Сер. 1: физика, математика и механика. —1988. — № 2. — С. 36—40.

21. *Воробьев О.Ю., Валендик Э.Н.* Вероятностное множественное моделирование распространения лесных пожаров. — Новосибирск: Наука, 1978. — 160 с.

22. *Раушенбах Г.В., Заславский А.А.* // Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях: Материалы I Всесоюз. школы-семинара. — Пущино: НЦБИ, 1986. — С. 126—141.

23. *Сердобольский В.И., Орлов А.И.* // Тез. докл. III Всесоюз. школы-семинара «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа». — М.: ЦЭМИ АН СССР, 1987. — С. 151—160.

24. *Орлов А.И.* Вероятностные модели объектов нечисловой природы // Заводская лаборатория. — 1995. — Т. 61, № 5.

25. *Орлов А.И.* Непараметрические оценки плотности в топологических пространствах // Прикладная статистика. — М.: Наука, 1983.— С. 12—40.

26. *Frascati* Manual: 1993. The Measurement of Scientific and Technological Activities. — Paris: OECD, 1994. — 261 с.

27. *Развитие* науки в России / ЦИСН. — М., 1993. — 468 с.

28. *Налимов В.В., Мульченко А.Б.* Наукометрия. — М.: Наука, 1969.

29. *Орлов А.И.* Прикладная статистика — «золушка» научно-технической революции // Наука и технология в России. — 1994. — № 1(3). — С.13—14.