1. Исходным рубежом интенсивного формирования информационного массива отечественных публикаций по Антарктике является 1997 год, что, безусловно, связано с началом исследований Украины на антарктической станции „Академик Вернадский” и в морских экспедициях.

2. В зарубежной библиографии по проблемам исследования Антарктики Украина представлена числом публикаций, которое в 2—3 раза меньше действительного их количества. Из работ украинских ученых в основном отражаются: доклады на международных конференциях; статьи в иностранных журналах; статьи в отечественных журналах, переиздающихся в переводе или распространяющихся по подписке за рубежом; статьи в отечественных журналах и сборниках, специально рассылаемых национальными операторами и авторами (напр. „Бюллетень...”).

3. Публикация в отечественных журналах статей на английском языке не дает никаких преимуществ с точки зрения быстроты распространения публикации за рубежом: определяющим фактором является не язык отдельных статей, а библиографическая доступность издания в целом.

Очевидно, задача состоит в том, чтобы от стихийного или выборочного пополнения отечественных и зарубежных баз данных о публикациях украинских полярных исследователей перейти к системному управлению формированием информационных потоков в этой области знаний.

Ряд мер, в основном организационного характера, т.е. не требующих особого финансирования, реально осуществить уже сейчас.

1. *Добров Г. М.* Наука о науке. — Киев: Наук. думка, 1970. — 320 с.

2. *Добров Г.М., Коренной А.А.* Наука: информация и управление. — М.: Сов. радио, 1977. — 256 с.

3. *Научно-технический* потенциал: структура, динамика, эффективность /Г.М.Добров, В.Е.Тонкаль, А.А.Савельев, Б.А.Малицкий. — Киев: Наук. думка, 1988. — 347 с.

Є.П.Удалов,
ст.наук.співроб., канд.фіз.-мат.наук (Київський національний університет ім.Т.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

Вступ

Г.М. Доброву був властивий надзвичайно широкий кругозор, що спирався на здатність справжнього наукового передбачення. Так, у своїй праці [1, с.282] видатний вчений зазначав (цитуючи мовою оригіналу): “...говоря об основной целевой установке науковедения, мы имеем в виду выработку научных основ оптимизации функционирования науки в целом. Подобного рода ответственные цели научного изучения опыта функционирования науки требуют, естественно, дальнейшего развития научно-методических

основ этой работы... Становление математического аппарата науковедения... включает в себя весьма широкий круг вопросов, решение которых может иметь далеко идущие последствия для повышения теоретического уровня и прикладной эффективности исследований в данной области.

Наряду с более широким и тонким использованием уже взятых на вооружение методов здесь предстоит освоить идеи и методы информатики, исследования операций, теории массового обслуживания, некоторые разделы теории множеств, экономико-математических методов и теории игр. На этой основе будет совершенствоваться системный подход к исследованию науки, развиваться ее структурный анализ и конкретно реализоваться идеи и возможности моделирования некоторых процессов научного развития (например, моделирование межнаучного взаимодействия).

Очевидно, что важнейшими предпосылками для успешного использования средств современной математики является становление специального языка науковедческих исследований, строгой системы понятий и, что особо важно, выработка системы конкретных измерителей для различных характеристик научного развития. Можно высказать убеждение, что исследования по этой проблеме науки о науке укрепят важное направление науковедения, заслуживающее названия «наукометрия»".

Такий прогноз Г.М. Доброва знаходить тверде підгрунтя як в історії наукових розробок у цих напрямках, так і в сучасних застосуваннях точних математичних методів для аналізу явищ у соціальній сфері.

Історично доведено, що класичною областю прикладної математичної статистики є математичні методи вибіркових досліджень [2—6]. Найбільше вони знаходять застосування у техніці, медицині, соціології. Починаючи з 1970—1975-х років у нашій країні розвиток сучасних вибіркових методів, зокрема статистики об'єктів нечислової природи, стимулювався запитами соціологічних і експертних досліджень [2]. О.І.Орловим зі співавторами розроблено нові підходи, сформульовано постановки, запропоновані алгоритми аналізу різнотипних даних (які включають значення кількісних і якісних ознак), отримано теореми про властивості цих алгоритмів, спроможність оцінок і т.п. Загалом зазначені теоретичні результати представлено у монографії [3].

Перехід до ринкової економіки в Україні й на теренах колишнього СРСР, що супроводжувався різким спадом виробництва, високим рівнем інфляції, дефіцитом державного бюджету, зменшенням попиту на дослідження і розробки з боку промисловості, негативно позначився на стані науки. Кризові процеси, що відбуваються у вітчизняній науці, взагалі потребують нових підходів як до методів збору даних про її стан (фінансування, результативність, матеріально-технічна база, кадри, інфраструктура і т.д.), так і до методів аналізу отриманих даних.

Дзеркалом процесів, які мають місце в останні роки в науці, можуть стати декілька 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. дол. США), але, враховуючи коєфіцієнти інфляції, спостережуване зростання не таке вже і велике. Тенденція у 2006 році подібна. Збереженню потенціалу української науки деякою мірою при-

діляється увага з боку міжнародної наукової громадськості, окремих країн, міжнародних організацій, зокрема Українського науково-технічного центру, загальне фінансування проектів яким у 2004 р. склало 17,9 млн. дол. США. Але провідні світові фірми вкладають у наукові розробки значно більші кошти: так, у 2005 р. фірма "Simens" — 6500 млн. дол. США, "Samsung" — 4500 млн. дол. США, "Microsoft" — 5800 млн. дол. США [7].

Коли йде мова про різноманітні задачі вивчення науки і керування нею, дуже важливими є вихідні методологічні принципи, однакові розуміння й оцінка обговорюваних процесів. Показовою є дискусія, в якій один з авторів як основний показник використовував продуктивність праці науковця [8], а інший — фондоємність наукової продукції [9], що явно ускладнювало взаєморозуміння. Суттєво, що в обох статтях широко використовувалися як статистичні, так і експертні дані. Тому беззаперечним є висновок, що саме статистичні дані про науковий потенціал — база для теоретичного і прикладного наукознавства [10,11].

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

З початку 70-х років минулого століття набула 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 вченим [12,13].

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

При аналізі даних про науковий потенціал методи статистики ОНП виявляються найбільш корисними, оскільки істотна частина даних має нечисловий, якісний (вербальний) характер [14—16].

Головним елементом математичної статистики є вибірка. Ймовірнісна теорія статистики показує, що вибірка — це сукупність незалежних, однаково розподілених випадкових 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 [7].

ОНП — це об'єкти, які недоцільно описувати числами, зокрема елементами нелінійних просторів. Прикладами є бінарні співвідношення, такі як розбивки, ранжировки, толерантності та ін., результати парних і множинних порівнянь, вимір у шкалах, відмінних від абсолютних, множини, нечіткі множини. Наведемо приклад ОНП, що мають якісні ознаки: стать людини чи тип наукової організації; взагалі результат віднесення об'єкта до однієї із заданих градацій (категорій); сукупність людей, які займаються визначеної працею (фізичною, розумовою та ін.); слово, пропозиція, текст, шрифт, котрі у пам'яті комп'ютера кодуються за допомогою цифр 0 і 1, але не стають від цього числами; розбивки об'єктів

на групи подібних між собою (кластери); ранжировки — 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 застосовується операція додавання (віднімання). Для розрахунку вибіркових характеристик розподілу таких математичних величин, як вибіркове середнє арифметичне, вибіркова дисперсія й т.п., у регресійному аналізі та інших областях цієї дисципліни постійно використовуються суми. Апарат математичної статистики оперує з великими числами, тому закони великих чисел, центральна гранична теорема та 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 вітчизняних вчених мають пріоритет на світовому рівні [13].

1.1. Середні дані, отримані теоретичним і практичним шляхом

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 як рішення екстремальних задач теоретичні та емпіричні середні [10]. Теоретичне середнє (у класичному розумінні) — це розв'язок задачі мінімізації математичного очікування відстані від випадкового елемента (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 вона мінімізується для цієї крапки. При цьому і емпіричне, і теоретичне середнє 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 вказаним способом, тобто збіжність емпіричних середніх до теоретичного приросту обсягу вибірки [3,18,19]. 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 задач прикладної статистики. Зокрема, крім законів великих чисел, встановлена і множина 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

Суттєвий інтерес становлять результати, пов'язані з конкретними областями статистики ОПН, зокрема зі статистикою нечітких множин, з випадковими множинами. Тут слід зазначити, що теорія нечітких множин у визначеному сенсі зводиться до теорії випадкових множин [3,13,20], до непараметричної теорії парних порівнянь з аксіоматичним введенням метрик у конкретних просторах ОПН [21].

Сучасні методи класифікації, у 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 [20—22].

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

1.3. Основи теорії вимірів

Розглянемо перехід від соціологічного завдання до математичного, а саме до однієї з наведених постановок проблеми однозначності в репрезентативній теорії виміру [5, 6]. Почнемо з розгляду конкретного соціологічного дослідження.

При вивченні привабливості різних професій для випускників шкіл [4] був складений список із 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 [4] використовувалося середнє арифметичне балів, виставлених професіям опитаними школярами. Зокрема, фізика одержала середній бал 7,69, а математика — 7,50. Відповідно до логіки [4] фізика краща, ніж математика.

Однак було відзначено [4], що цей висновок суперечить даним праці [23], згідно з якими школярі середніх класів більше люблять математику, ніж фізику. Обговоримо одне з можливих пояснень цього протиріччя, що полягає в неоднаковій методиці обробки даних, застосованих у праці [4].

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 зовсім не випливає, 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 за привабливістю, властиве школярам, не залежить від того, якою системою балів їм запропонує користуватися соціолог. Коли так, то розподіл професій за градацією 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 пере-

ходу до нової системи балів відповіді щодо математики будуть мати вигляд $\gamma(x_1), \gamma(x_2),... \gamma(x_n)$, а щодо фізики — $\gamma(y_1), \gamma(y_2),... \gamma(y_n)$.

Нехай єдина оцінка привабливості професії обчислюється за допомогою функції $f(x_1, x_2,...x_n)$. Які вимоги природно накласти на функцію $f: K^n \to K^1$, щоб отримані з її допомогою висновки не залежали від того, якою саме системою балів користувався соціолог ?

Єдина оцінка обчислювалася для того, щоб порівнювати професії за привабливістю. Тому зажадаємо стійкості результату порівняння [23]: нерівність

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

справедлива тоді й тільки тоді, коли справедлива нерівність

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

причому однозначність нерівностей (1) і (2) є при будь-яких x_1, y_1 і γ. Які f стійкі щодо порівняння? Відповідь на це питання було дано у праці [24]. Зокрема, з'ясувалось, що середнім арифметичним, як у праці [4], користуватися не можна, а членами варіаційного ряду (і тільки ними) — можна і необхідно.

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: одновимірна 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 ранжировок і n толерантностей, або n множин, або n нечітких множин і т.п. [22, 25, 26].

Тому відзначимо необхідність розвитку методів статистичної обробки "різнотипових даних", обумовлену великою роллю в прикладних дослідженнях "ознак змішаної природи" [27].

Результат спостереження стану об'єкта найчастіше являє собою вектор, в якого частина координат вимірюється 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 не придатні для обробки різнотипних даних. Є й більш складні моделі різнотипових даних, наприклад, коли деякі координати вектора спостережень описуються нечіткими множинами [28].

Для позначення подібних некласичних результатів спостережень [27] запропонований збірний термін — ОНП (об'єкти нечислової природи). Термін "нечисловий" [29] означає, що структура простору, 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 перерахо-

ваних 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 [30].

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 теорем, що доводяться).

Інакше кажучи, у цьому випадку передбачається, що результати спостережень, як правило, це елементи вибірки, і вони лежать у довільному просторі $\mathcal{X}$. Для одержання теорем необхідно зажадати, щоб $\mathcal{X}$ задовольняло деяким умовам, наприклад було топологічним простором [31]. 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 наші 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. Мета 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 [19—21]. Погляд із загальної точки зору дозволяє одержати нові результати й в інших областях математичної статистики [22, 25, 26, 32].

1.5. Використання вербальних об'єктів при формуванні математичної моделі

Використання ОНП часто породжене бажанням обробляти об'єктивнішу, більш звільнену від похибок інформацію. Як показали численні досліди, людина правильніше (і з меншими утрудненнями) відповідає на питання якісного, порівняльного характеру, ніж кількісного. Так, їй легше сказати, яка із двох гир важча, ніж указати їх приблизну 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 [3] при аналізі частинних 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 [2].

Науку про єдність мір і точність вимірів називають метрологією. Таким чином, репрезентативна теорія вимірів — частина метрології. Методи обробки даних повинні бути адекватні щодо припустимих перетворень шкал виміру в сенсі репрезентативної теорії вимірів [33]. Однак встановлення типу шкали, тобто задання групи перетворень Ф — справа фахівця відповідної прикладної області. Так, оцінки привабливості професій вважались вимірюваними в порядковій шкалі [3]. Проте окремі соціологи не погоджува-

лися із цим, вважаючи, що випускники шкіл користуються шкалою з більш вузькою групою припустимих перетворень, наприклад інтервальною. Очевидно, ця проблема стосується не математики, а наук про людину [11, 26, 32—35]. Для її розв'язання 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 зазначали, номінальні й порядкові шкали широко поширені не тільки в соціально-економічних дослідженнях. Вони застосовуються в медицині, мінералогії, географії й т.д. Нагадаємо, що за шкалою інтервалів вимірюють величину потенційної енергії або координату крапки на прямій, на якій не відзначені ні початок, ні одиниця виміру; за шкалою відношень — більшість фізичних одиниць: масу тіла, довжину, заряд, а також ціни в економіці. Час вимірюється за шкалою різниць, якщо рік приймаємо природною одиницею виміру, і за шкалою інтервалів — у загальному випадку. У процесі розвитку відповідної області знання тип шкали може змінюватися. Так, спочатку температура вимірювалась за порядковою шкалою (холодніше — тепліше), потім — за інтервальною (шкали Цельсія, Фаренгейта, Реомюра) і, нарешті, після відкриття абсолютного нуля температур — за шкалою відношень (шкала Кельвіна). Слід зазначити, що серед фахівців іноді є розбіжності з приводу того, за якими шкалами та репрезентативними вибірками варто вимірювати ті або інші реальні величини.

Відзначимо, що термін "репрезентативна" використовується, щоб відрізнити розглянутий підхід до теорії вимірів від класичної метрології, а також від робіт А.М.Колмогорова і 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 звернемо увагу на те, що аналізовані в класичній статистиці результати спостережень також "не зовсім числа". А саме будь-яка величина X вимірюється завжди з деякою похибкою ΔX, і результатом спостереження є

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

Як ми вже зазначали, похибками вимірів займається метрологія причому має місце наступне:

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

б) розподіл ΔX у переважній більшості випадків не є нормальним [3];

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

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

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

Строго формулюючи, 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 неперебoрною похибкою ΔX. Перший підхід на сьогодні недоцільний, тому що він вимагає відмови від більшої частини розробленого математичного апарату. Із другого випливає необхідність вивчення впливу неперебoрних похибок на статистичні висновки.

Похибки ΔX можна враховувати або за допомогою ймовірнісної моделі (ΔX — випадкова величина, яка має функцію розподілу, загалом кажучи, що залежить від X), або за допомогою нечітких множин. У другому випадку приходимо до теорії нечітких чисел і до її часткового випадку — статистики інтервальних даних [16,17].

Інше джерело появи похибки ΔX пов'язане із прийнятою в конструкторській і технологічній документації системою допусків на контрольовані параметри виробів і деталей, з використанням шаблонів при перевірці контролю якості продукції [16]. У цих випадках характеристики ΔX визначаються не властивостями засобів виміру, а застосовуваною технологією проектування й виробництва. У термінах прикладної статистики сказаному відповідає групування даних, коли ми знаємо, до якого із заданих інтервалів належить спостереження, але не знаємо точного значення результату спостереження. Застосування групування може дати економічний ефект, оскільки найчастіше легше (у 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 у процесі обробки даних і на "виході" як підсумок статистичного аналізу.

Розглянемо найпростішу прикладну постановку завдання регресії [3]. Вихідні дані мають вигляд $(x_i,\ x_i) \in \mathrm{R}^2$, $i=1,2,...,n$. Ціль полягає в тому, щоб із достатньою точністю описати y як багаточлен (поліном) від x, тобто модель має вигляд

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

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

Тут наочно проявляється одна з причин живучості статистичних моделей на основі нормального розподілу. Такі моделі, хоча й, як правило, не адекватні реальній ситуації [8, 9], 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 спрямовані на зняття нереалістичного припущення нормальності й перехід до непараметричних моделей похибок.

Розповсюджена процедура відновлення залежності за допомогою багаточлена 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 [3]. 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), множину значень якого можна описати як $R^1 \bigcup R^2 \bigcup R^3 \bigcup ...$ Це об'єкт нечислової природи, звичайні методи оцінювання для нього незастосовні, тому що m — не дискретний параметр. У розглянутій постановці розроблені до теперішнього часу методи оцінювання ступеня полінома мають в основному евристичний характер [17]. Властивості описаної вище розповсюдженої процедури розглянуті в [3]. Показано, що методами, якими звичайно користуються, ступінь полінома m оцінюється недостовірно, тому знайдено граничний розподіл оцінок цього параметра, який виявився геометричним. Відзначимо, що для ступеня багаточлена раніше запропоновані достатні оцінки [14].

У більш загальному випадку лінійної регресії дані мають вигляд (y_i, X_i), $i = 1,2,...,n$, де $X_i = (x_{i1}, x_{i2}, ..., x_{iN}) \in R^N$ — вектор предикторів (факторів, що пояснюють змінні), а модель така:

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

(тут 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

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

У моделі (4) є природний порядок введення предикторів у розгляд — відповідно до зростання ступеня, а в моделі (5) природного порядку немає, тому тут стоїть довільна підмножина множини предикторів. Існує тільки частковий порядок: чим потужність підмножини менше, тим краще. Модель (5) особливо актуальна в технічних дослідженнях [14, 16—18, 33]. Вона застосовується в задачах керування якістю продукції й інших техніко-економічних дослідженнях, у медицині, економіці, маркетингу й соціології, коли з великої кількості факторів, що наближено впливають на досліджувану змінну, треба відібрати по можливості найменшу кількість 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 розв'язання 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 відзначали, існують методи, в основному евристичні, які найчастіше не є навіть обґрунтованими. Тому 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 — його оцінка. Оцінка

K_n називається обґрунтованою, якщо

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

де Δ — символ симетричної різниці множин; Card(K) означає число елементів множини K, а границя 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 моделі (4) структура описується невід'ємним 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 вивчене, 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 в багатовимірному шкалуванні, в інших оптимізаційних постановках проблем прикладного багатовимірного статистичного аналізу [4—6, 19—26].

Перейдемо до об'єктів нечислової природи на "виході" статистичної процедури. Приклади численні. Розбивки — підсумок роботи багатьох алгоритмів класифікації, зокрема алгоритмів кластер-аналізу. Ранжировки — результат упорядкування професій за привабливістю або автоматизованої обробки думок експертів — членів комісії з підведення підсумків конкурсу наукових праць (в останньому випадку використовуються ранжировки зі зв'язками: так, в одну групу, найбільш численну, попадають роботи, що не одержали нагород). Із усіх об'єктів нечислової природи, мабуть, найбільш часті на "виході" дихотомічні дані [3]: прийняти або не прийняти гіпотезу, прийняти або забракувати партію продукції. Результатом статистичної обробки даних може бути множина, наприклад зона найбільшого ураження при аварії, або послідовність множин, наприклад "середньовимірний" опис поширення пожежі [28]. Нечіткою множиною E Борель [21] ще на початку XX в. 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 механічних систем, 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

Через цілий ряд причин — високу динаміку змін, потребу в оперативній інформації для прийняття рішень і т.п. — виникає гостра необхідність регулярного проведення оперативних обстежень наукових організацій і соціологічних опитувань вчених [7—9, 11, 32, 34—36]. 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 наукового потенціалу і керування ним.

Вивчення показників наукової діяльності та експертно-статистичний підхід до побудови інтегральних показників. Статистичний і експертний аналіз показників науки, виявлення їх рейтингу: які показники є найбільш значимими із врахуванням загальноприйнятих міжнародних стандартів у статистиці науки? 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 занадто складна, то чи не вдасться її спростити, розглядаючи наукові організації з визначених практичних позицій — з погляду наукового рівня, віддачі, фінансування, перспективності та виживання і т.д ?

Також становить інтерес структура наукових рад, комісій та інших форм діяльності вчених, особливо в ситуації, коли через такі структури надходить фінансування.

Рейтинги наукових 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 пророблення. Можна вказати кілька підходів. Для розв'язання цієї задачі можна використовувати три варіанти експертно-статистичного методу інтегральної оцінки факторів перспективності наукового пошуку або напрямку наукової організації:

- *за інтегральним показником (лінійна функція, параметри якої — показники науки, а значення коефіцієнтів надходять від експертів);*
- *за навчальними вибірками, отриманими від кваліфікованих експертів;*
- *за допомогою оцінки параметрів інтегрального показника згідно з навчальними вибірками (власне експертно-статистичний метод).*

Розрахунок і короткострокове прогнозування показників 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 урахуванням 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 та інтервальних даних. Використання статистики об'єктів нечислової природи у вибіркових обстеженнях, зокрема регресійного аналізу у просторах різнотипових ознак (об'єктів нечислової природи), дасть можливість оцінити ефективність фінансування, а статистика інтерваль-

них даних дозволить врахувати неминучі похибки у наявних даних.

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

Прогноз "виживання" НДІ може бути побудований за допомогою самонавчальних вибірок на основі непараметричних оцінок щільності в просторі різнотипових ознак, частина координат яких — кількісні ознаки, а частина — якісні.

Існує багато інших цікавих проблем [11, 34]. Наприклад, виявлення циклічності розвитку науково-технічного потенціалу країни, вивчення динаміки реальної і формальної структури науки [35], форм примітивізації наукової діяльності й наукової продукції в умовах різкого спаду виробництва і скорочення фінансування наукової праці, проблеми відображення в суспільній свідомості та у самосвідомості наукового співтовариства специфіки наукової діяльності (включаючи аналіз розповсюджених догм) та ін.

За нашою оцінкою, застосування сучасних статистичних методів у вибіркових дослідженнях наукових організацій дозволить одержувати результати, цікаві з теоретичної точки зору і корисні для практики керуючих рішень, спрямованих на розвиток науки.

1. *Добров Г.М.* Наука о науке: Начала науковедения. — 3-е изд., доп. и перераб. — Киев: Наук. думка, 1989. — 301 с.

2. *Орлов А.И.* Нечисловая статистика // Наука и технология в России. — 1994. — № 3(5). — С.7—8.

3. *Анализ* нечисловой информации в социологических исследованиях / Под ред. В. Г. Андреенкова, А. И. Орлова, Ю. Н. Толстовой. — М.: Наука, 1985. — 220 с.

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

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

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

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

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

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

10. *Орлов А.И.* Допустимые средние в некоторых задачах экспертных оценок и агрегирования показателей качества // Многомерный статистический анализ в социально-экономических исследованиях. — М.: Наука, 1974. — С.388—393.

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

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

13. *Орлов А.И.* Статистика объектов нечисловой природы и экспертные оценки // Экспертные оценки. — М.: Научный совет АН СССР по комплексной проблеме "Кибернетика", 1979. — С.17—33. — (Вопросы кибернетики; Вып.58).

14. *Орлов А.И.* Статистика объектов нечисловой природы: Обзор // Зав. лаб. — 1990. — Т.56, № 3. — С.76—83.

15. *Orlov A.I.* On the Development of the Statistics of Nonnumerical Objects // Design of Experiments and Data Analysis: New Trends And Results / Ed. by prof.E.K.Letzky. — Moscow: ANTAL, 1993. — P.52—90.

16. *Орлов А.И.* Объекты нечисловой природы // Зав. лаб. — 1995. — Т.61, № 3.

17. *Орлов А.И.* Вероятностные модели объектов нечисловой природы // Там же. — 1995. — Т.61, № 5.

18. *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).

19. *Орлов А.И.* Асимптотическое поведение статистик интегрального типа // Вероятностные процессы и их приложения. — М.: МИЭМ, 1989. — С.118—123.

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

21. *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.

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

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

24. *Карапетян К.А., Чахмахчян 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.

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

26. *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.

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

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

29. *Раушенбах Г.В., Заславский А.А.* // Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях: Материалы 1-й Всесоюз. школы-семинара. — Пущино: НЦБИ, 1986. — С. 126—141.

30. *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.

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

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

33. *Орлов А.И.* Связь между средними величинами и допустимыми преобразованиями шкалы // Математические заметки. — 1981. — Т.30, № 4. — С.361—368.

34. *Развитие* науки в России. — М.: ЦИСН. — 1993. — 468 с.

35. *Орлов А.И.* Прикладная статистика — "Золушка" научно-технической революции // Наука и технология в России. — 1994. — № 1(3). — С.13—14.

36. *Дело.* — 2005. — 10 ноября (№ 17). — С.12.