Окремі випадки застосування методів чисельного ресамплінгу в прикладних задачах

М. НАКОПІЯ

Дніпродзержинський державний технічний університет

У даній роботі досліджувались методи чисельного ресамплінгу, які об'єднують три різних підходи, що відрізняються за алгоритмом, але близькі по суті: рандомізації (або перестановочний тест), бутстрапу і метод «складного ножа». Також розглянуті області їх використання, переваги і недоліки. За допомогою програмної реалізації здійснено порівняльний аналіз варіації бутстрапу.

В данной работе исследовались методы численного ресамплинга, которые объединяют три разных подхода, отличающихся по алгоритму, но близких по сути: рандомизации (или перестановочный тест), бутстрапа и метод «складного ножа». Также рассмотрены области их применения, преимущества и недостатки. С помощью программной реализации осуществлен сравнительный анализ вариаций бутстрапа.

In this paper we investigated methods of numerical resampling, which combines three different approaches, which differ according to the algorithm, but similar in essence: randomization (or permutation), bootstrap and jackknife method. Also considered fields of application, advantages and disadvantages. With the help of a software realization, the comparative analysis of variations bootstrap has been implemented.

Вступ. Стрімка зміна сучасного світу, пов'язана з революційними досягненнями обчислювальної техніки, інформаційних технологій і зв'язку, забезпечила можливість швидкого, комплексного і точного аналізу великих масивів даних.

Менш очевидний процес пов'язаний з докорінним переглядом основних концепцій прикладної статистики. У докомпьютерний період, коли обробка даних вимагала багато часу і зусиль, робився акцент на методи, які дозволили б отримати максимум інформації при невеликому обсягу обчислень. Загальний підхід був вельми простий: робилося припущення, що структура отриманих даних "схожа" на деяку поширену статистичну модель (наприклад, підпорядковується нормальному розподілу), після чого вибіркові оцінки параметрів розраховувалися по відносно простим теоретичним формулам. Однак для складних систем (перш за все, економічних і екологічних), що складаються з великої кількості неоднорідних компонент, в структурі даних спостерігається істотна відміна від звичайних гаусівських розподілів. Зокрема, феномен полягає в тому, що в результаті збільшення обсягу вибірки деякі оцінювані параметри генеральної сукупності (в першу чергу, дисперсія) починають монотонно зростати, тобто дані перестають підкорюватися центральній граничній теоремі теорії ймовірностей [1]. У цих випадках висновки, засновані на припущеннях про нормальність, часто не є коректними і тому практично завжди виявляються некорисними.

Поява комп'ютерів докорінно змінила концепцію обробки даних, так як обчислення стали швидкі і необтяжливі, але постала вимога коректності сформованих висновків. Відомий американський статистик, професор Стенфордського університету Б. Ефрон написав статтю під назвою «Комп'ютери та статистика: подумаємо про неймовірне» [2], в якій обгрунтував розвиток нового класу альтернативних комп'ютерно-інтенсивних (computer-intensive) технологій, що включають рандомізацію, бутстрап і методи Монте-Карло. Ці технології, об'єднані загальним терміном "чисельний ресамплінг", не вимагають ніякої апріорної інформації про закон розподілу досліджуваної випадкової величини, замість цього вони виконують багаторазову обробку різних фрагментів вихідного масиву емпіричних даних, ніби розглядаючи їх з різних боків і порівнюючи отримані таким чином результати.

З огляду на це можна припустити, що розвиток прикладної статистики піде по двох різних шляхах. Перший полягає в розвитку традиційного "асимптотичного" напряму і в його рамках розширюється арсенал методик і нових критеріїв, які можуть виявитися кращими в тих чи інших умовах обробки даних. Але, наприклад, в ході дисперсійного аналізу при різних його модифікаціях рекомендовано використовувати близько трьох десятків "іменних" критеріїв (Дана, Коновер, Джонкхіера-Терпстра, Бартлетта, Кокрена, Шеффе, Дункана, Тьюки, Льовені, Кульбака та ін.). Для перевірки нормальності розподілу - більше двох десятків критеріїв згоди, а в непараметричній статистиці число методик порівняння вибірок, представлених в довідниках [3], наближається до сорока.

Області використання кожного з цих варіантів виглядають розмитими, а переваги і недоліки, що відзначаються суб'єктивні і суперечливі. Альтернативний шлях зводиться до розробки єдиних універсальних алгоритмів пошуку рішення (наприклад, формування частотного розподілу аналізованого показника в результаті багаторазових ітерацій). Це дозволяє тільки за рахунок інтенсивної роботи комп'ютера провести надійне тестування даних без суворої прив'язки до формули, яка застосовується критерієм. Так як статистика заснована на обчисленнях, ефективність і результативність їх реалізації повинна бути найбільш важливим і об'єктивним аргументом у вирішенні, який з цих двох шляхів обробки даних краще підходить для широкого кола прикладних задач.

Ресамплінг ґрунтується на традиційних загальних ідеях статистичного аналізу. Фундаментальним залишається міркування про співвідношення між випадковими повтореннями емпіричних даних і генеральною сукупністю. Статистичні висновки також базуються на класичних інтервалах надійності і p-значеннях, заснованих на вибіркових розподілах використовуваних критеріїв. Ключова відмінність лише в тому, що повторності класичної вибірки витягуються з генеральної сукупності, а псевдо повторності ресамплінгу - з самої

емпіричної вибірки [4].

При цьому, по-перше, нові методи ресамплінгу звільняють нас від необхідності робити не завжди обґрунтовані припущення. По-друге, вони безпосередньо звертаються до самої суті статистичного аналізу і показують, як зміниться розподіл вибіркових характеристик, якщо буде використано практично необмежену кількість повторностей даних, отриманих в тих самих умовах. По-третє, при достатній кількості проведених ітерацій методи ресамплінгу дають більш точні результати, ніж традиційні методи. Нарешті, вони концептуально простіші і звільняють нас від необхідності шукати в довідниках різні математичні формули критеріїв, найбільш придатних в конкретних умовах, і способів їх апроксимації.

У довгостроковій перспективі перераховані переваги ресамплінгу можуть мати вплив на стиль, з яким предмет статистики викладається і здійснюється на практиці.

Хоча самі процедури рандомізації і бутстапа концептуально дуже прості, головна причина недостатнього практичного використання цих методик розрахунку зазвичай пояснюється відсутністю необхідного програмного забезпечення. Взагальному випадку для реалізації такої можливості є два підходи:

- використання програм, керованих за допомогою меню, таких як SPSS або пакет Statistica;

- запис макровизначень на високорівневих інтерпретованих мовах в обчислювальних інтерактивних середовищах, таких як R, MatLab, SAS, Stats, або самостійна розробка програм з використанням VisualBasic, C ++, Delphi та інших.

Методи ресамплінгу об'єднують три різних підходи, що відрізняються за алгоритмом, але близькі по суті: рандомізацію, або перестановочний тест (permutation), бутстрап (bootstrap) і метод «складного ножа» (jackknife).

Ідеї чисельного ресамплінгу не є принципово новими в статистиці і відносяться принаймні до 1935 року, але практичне застосування цих методик було пов'язано з вимушеним очікуванням, поки не з'являться досить швидкі комп'ютери. Ідея методу «складного ножа» (або jackknife) полягає в тому, щоб послідовно і багаторазово виключати з наявної вибірки, яка налічує n елементів, по одному її члену і обробляти варіаційний ряд з решти $(n-1)$ елементів [5]. Середнє значення або медіана буде при цьому «блукати» і тоді можна проаналізувати інформацію при кожному зміщенні, побудувати розподіл вибіркової оцінки параметра, який ми шукаємо і уточнити його властивості. Ф. Мостеллер і Дж. Тьюки [6] вважали цей алгоритм «універсальною методикою».

Бутстрап-процедура (або bootstrap) була запропонована як узагальнення алгоритму «складного ножа», щоб не зменшувати кожен раз число елементів в порівнянні з вихідною сукупністю. Основна ідея бутстрапа полягає в тому, що методом статистичних випробувань Монте-Карло багаторазово витягувати повторні вибірки з емпіричного розподілу. А саме: беруть кінцеву сукупність з n членів вихідної вибірки $x_1, x_2, ..x_{n-1}, x_n$ звідки на кожному кроці з n послідовних ітерацій за допомогою датчика випадкових чисел, рівномірно розподілених на інтервалі $[1, n]$ «витягується» довільний елемент x_k, який знову «повертається» в вихідну вибірку (може бути витягнутий повторно). Наприклад, при $n = 8$ одна з таких комбінацій має вигляд $x_4, x_2, x_8, x_2, x_1, x_2, x_5, x_4$ тобто окремі елементи можуть повторюватися. Цим способом можна сформувати будь-яке велике число бутстрап-вибірок. Як і у разі «складного ножа», в результаті легкої модифікації частотного розподілу реалізацій вихідних даних можна очікувати, що кожна наступна псевдовибірка, яка генерується буде повертати значення параметра, який трохи відрізняється від обчисленого для початкової сукупності. Утворений розкид значень показника дає можливість побудови надійних інтервалів і інших корисних вибіркових параметрів аналізованої величини [7].

Одночасно з впровадженням методів планування експерименту почали бурхливо розвиватися алгоритми рандомізації, які полягають в багаторазовому випадковому перемішуванні рядків або стовпців таблиці спостережень щодо рівнів впливу досліджуваних факторів. При кожній ітерації перестановочного тесту на основі згенерованої псевдо вибірки розраховуються імітовані значення t_{TAN} аналізованого показника або статистики, які порівнюються з аналогічною величиною t_{obs}, знайденої за емпіричними даними. В ході перестановок не змінюється ні склад вихідної таблиці, ні чисельність груп з різними рівнями впливу, а лише відбувається безпорядний обмін елементами даних між цими групами.

Існують думки, що рандомізація взагалі є окремим випадком випробувань Монте-Карло [8]. Однак незважаючи на схожість цих методів в принципових алгоритмів і обмеженнях, між ними є досить істотні концептуальні відмінності. Наприклад, для методів Монте-Карло типові дослідження, коли дані спостережень взагалі не використовуються, щоб змоделювати імовірнісний процес.

Процедури ресамплінгу не вимагають ніякої апріорної інформації про закон розподілу досліджуваної випадкової величини і в цьому сенсі можуть розглядатися як непараметричні. Вони виконують обробку різних фрагментів вихідного масиву емпіричних даних, як би повертаючи їх «різними гранями» і зіставляючи отримані таким чином результати. Питання про повну коректність такого прийому залишається відкритим, але якщо визнати його законним, то асимптотичні переваги ресамплінгу в порівнянні з класичними параметричними тестами стають очевидними. Значення параметрів, побудованих по розмноженим підвибіркам не є незалежними, однак при збільшенні n з ресамплірованими значеннями статистик можна звертатися як з незалежними випадковими величинами.

Постановка задачі. Спочатку розглянемо бутстрап-метод та його варіації. Бутстрап був введений Ефроном для незалежних однаково розподілених даних. Даний метод дозволяє отримати більш точне наближення розподілу статистики, що цікавить нас, ніж при використані асимптотичного розподілу. Але у випадку часових рядів звичайний бутстрап непридатний, тому що він порушує структуру часового ряду. Проте існують модифікації, які дозволяють застосувати первісну ідею в контексті тимчасових рядів. Різні варіації бутстрапу для часових рядів можуть бути представлені наступними рис. 1.1.

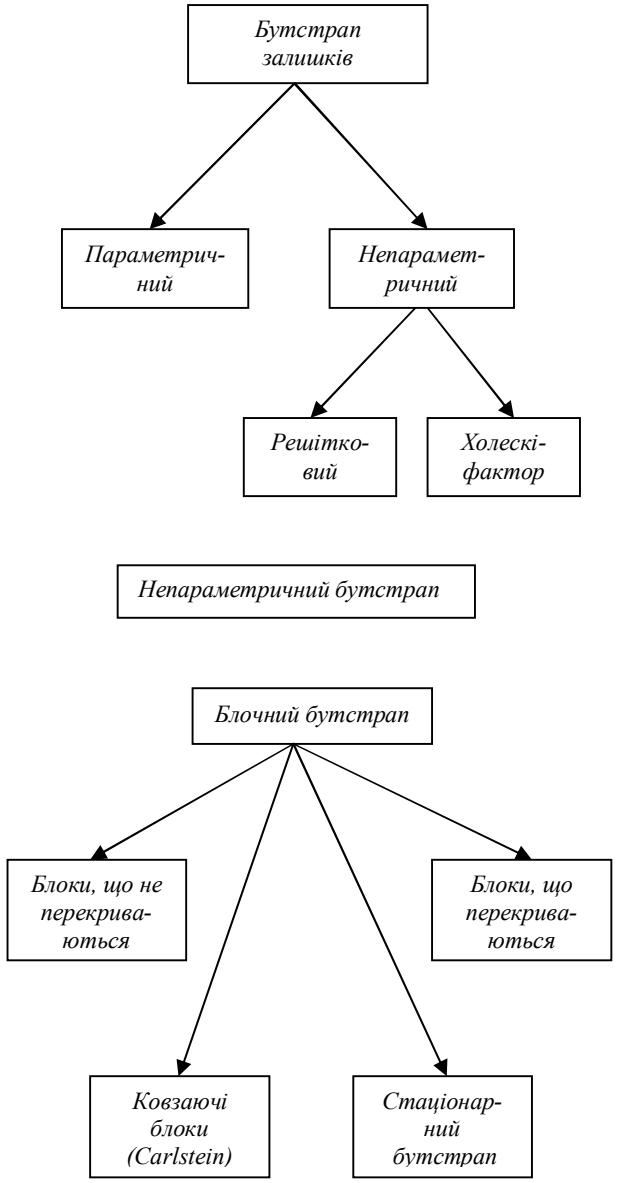

Рис. 1.1. Варіації бутстрапу

Серед багатьох варіацій бутстрапу, розглянутих раніше, зупинимось на практичному використані марковського бутстрапу, що за дослідженнями проведеними різними авторами, є достатньо ефективним при аналізі часових рядів.

Загальна постановка. Маємо вихідну вибірку $y_1,..y_T$ і побудовану частину псевдовибірки $y_1^*,..y_S^*$. Тоді $y_{S+1}^* = y_t$ з ймовірністю p_t:

$$p_t = p_t(y_1^*,..y_S^*; y_1,..y_T; S).$$

Необхідно знайти ці ймовірності, виходячи з наближення вихідного процесу дискретним марковським ланцюгом, який має в якості станів спостереження вихідної вибірки, а перехідні ймовірності знаходяться по деякому правилу і залежать від вихідної вибірки.

Найпростіший випадок марковського бутстрапу. Розглянемо пари $\{(y_k, y_{k+1})\}_{k=1}^{T-1}$. Якщо розглянути ці точки в двовимірної площині, то ці точки знаходяться в квадраті $[y_{(1)}, y_{(T)}]\times[y_{(1)}, y_{(T)}]$. Розділимо цей квадрат на частини, для цього поділимо кожну сторону на I частин $\Delta_1, \Delta_2, ..\Delta_I$ можливі варіанти такого розбиття будуть дані нижче. Після розбиття знайдемо відрізок Δ_k в який потрапила y_S^*, після чого визначимо ймовірності:

$$p_t = \begin{cases} 0, y_{t-1} \notin \Delta_k, \\ 1/\#\{y_1 \in \Delta_k\}, y_{t-1} \in \Delta_k \end{cases}$$

Тобто, якщо y_S^* потрапила в Δ_k, то ми з однаковою ймовірністю переходимо в точки, у яких передісторія також потрапляє в цей відрізок.

Вибір розбиття $[y_{(1)}, y_{(T)}]$ на відрізки Δ_i є одним з параметрів моделі. У симуляціях були розглянуті рівномірний емпіричне розбиття і розбиття на рівні відрізки. Позначимо $\Delta_i = [a_i, a_{i+1}]$.

Рівномірне емпіричне розбиття (з однаковою кількістю точок передісторії в кожному відрізку):

$$a_1 = y_{\min}, \quad \#\{a_i \le y_S \le a_{i+1}\} = m, \quad i = 1,...I,$$

$$I = \frac{\mathrm{T}}{\mathrm{m}}, \ (m = pT).$$

Найпростіший випадок згладженого бутстрапу. Розглянемо випадок переходу псевдовибірки на відрізок Δ_k, тоді ймовірності переходу в точку вихідної вибірки, у яких передісторія також потрапила в цей відрізок, будуть однаковими. Природним узагальненням є врахування того, наскільки далеко лягла передісторія у вибірці від передісторії точки, в яку ми хочемо перейти. Цю інформацію можна врахувати за допомогою ядерної функції, тоді отримуємо наступну формулу для перехідник ймовірностей:

$$p_t = p_t(y_1^*,..y_S^*; y_1,..y_T; S) = K_{h_{(y_S^*)}}(y_S^* - y_t),$$

де в якості ядерної функції K можливі різні варіанти, а також $h(y_S^*)$ є окремим параметром вибору. Ця формула нагадує непараметричний бутстрап Хансена в простому випадку (без використання обмежень на моменти), але тут h залежить від y_S^*, що дає додаткові переваги. як показали розрахунки, при h які не залежать від передісторії, метод дає гірші результати.

Результати. Отримано результати програмної реалізації результатів досліджуваних методів, які показали, що псевдовибірка, побудована за допомогою алгоритму марковського бутстрапу, дає краще наближення, ніж згладжений бутстрап.

	Параметры исходной выборки	Параметры псевдовыборки
Мат. ожидание	2,02466	2,13486
Дисперсия	1,080921	1,099316
Ср. квадратическое отклонение	1,039674	1,048483

Рис.1.2. Параметри вихідної вибірки та псевдовибірки, побудованої марковським бутстрапом

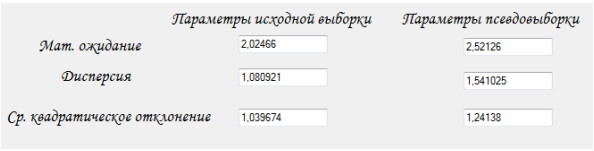

Рис.1.3. Параметри вихідної вибірки та псевдовибірки, побудованої згладженим бутстрапом.

Висновки.

В роботі досліджувались актуальні методи прогнозування. З'ясовувалось в яких випадках доцільно застосовувати ресамплінг, а в яких - інші звичайні статистичні методи.

Підкреслено фундаментальну відмінність між рандомізацією і бутстрапом: якщо рандомізаційний тест застосовується, щоб оцінити ступінь впорядкованості структури даних або взаємозв'язки між окремими її фрагментами, то бутстрап, або «складаний ніж», використовується для отримання найбільш коректної оцінки параметрів розподілу випадкової величини (середнього, медіани, дисперсії і т. д.).

Розглянуто бутстрап-метод та його варіації. За допомогою програмної реалізації зроблено висновки, що псевдовибірка, побудована за допомогою алгоритму марковського бутстрапу, дає краще наближення, ніж згладжений бутстрап.

Використання методів чисельного ресамплінгу є досить актуальним у розрізі застосування до великої кількості прикладних задач.

ЛІТЕРАТУРА

1. Хайтун С.Д. Негауссовость социальных явлений // Социологические исследования. 1983. № 1. С. 144-152.
2. Efron B., Tibshirani R.J. An intrduction to the bootstrap. N. Y.: Chapman &Hall, 1993. 436 p.
3. Гайдышев И. Анализ и обработка данных: специальный справочник. СПб: Питер, 2001. 7
4. Fox J. An R and S-Plus Companion to Applied Regression. London: Sage Publications Inc., 2002. 328 p.
5. Tukey J.W. Bias and confidence in not quite large samples // Ann. Math. Statist. 1958. V. 29. P. 614.
6. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. М:Финансы и статистика, 1982. Вып. 1. 320с.
7. Manly B.F.J. Randomization, bootstrap and Monte Carlo methods in biology. London:Chapman & Hall, 2007. 445 p
8. Efron B. Computers and the theory of statistics: thinking the unthinkable // SIAM Review. 1979a. V. 21, № 4. P. 460-480.
9. Анатольев С. Основы бутстрапирования //Квантиль. 2007. №3. С. 1-12.
10. Efron B. Bootstrep methods. Another look at the Jacknife // Ann. Statist. 1979. №7. P. 1-26.

пост. 24.02.2016

Використання методу експертних оцінок Дельфі у задачах прийняття рішень

Н.О. КУПІЧ

Дніпродзержинський державний технічний університет

Розглянуто один з методів експертних оцінок, який дозволяє на основі отриманих чи існуючих даних прогнозувати поведінку досліджуваної області на досить великі проміжки часу. Приведено приклад використання даного методу для вирішення прикладної задачі.

Рассмотрен один из методов экспертных оценок, который позволяет на основе полученных или уже существующих данных прогнозировать поведение изучаемой области на достаточно большие промежутки времени. Приведен пример использования данного метода для решения прикладной задачи.

Considered one of the methods of peer review, which allows on the basis of the existing data or predict the behavior of the study area for quite long periods of time. Also given an example of using this method to solve a practical problem.

Вступ. Метод Дельфі є найбільш формальним з усіх методів експертного прогнозування і частіше за інші використовується в технологічному прогнозуванні, дані якого використовуються потім у плануванні виробництва та збуту продукції. Це груповий метод, при якому проводиться індивідуальне опитування групи експертів щодо їх припущень про майбутні події в різних областях, де очікуються нові відкриття або вдосконалення. Таким чином, є можливість прогнозування розвитку тієї чи іншої сфери від 5 до 10 років.

В багатьох країнах виділяють такі теми для прогнозування :

- матеріали та їх обробка;
- інформатика;
- електроніка;
- охорона здоров'я та соціальне забезпечення,
- вивчення і використання космічного простору,
- енергетика та природні ресурси,
- екологія, сільське господарство,
- промислове виробництво,
- урбанізація і будівництво,
- зв'язок,
- транспорт [1].

Даний підхід полягає в опитуванні, яке проводиться в два або більше кіл. Опитування проводиться за допомогою спеціальних анкет анонімно, тобто особисті