

ПРОБЛЕМИ ОБРОБКИ РЕЗУЛЬТАТІВ СТАТИСТИЧНИХ СПОСТЕРЕЖЕНЬ

СЕРОВА І. А.

кандидат економічних наук

Харків

Першочерговою умовою проведення будь-якого статистичного дослідження є отримання інформації щодо його об'єкта, а також узгодженість мети й вимог щодо результатів його здійснення, котрі в подальшому визначають взаємозв'язок методів статистичного аналізу та процесу організації збирання даних. Отримання вихідних даних для статистичного дослідження може бути здійснено як через експеримент, так і через статистичне спостереження.

Якщо експеримент є методом вивчення об'єкта, коли дослідник активно й цілеспрямовано впливає на нього шляхом створення штучних умов (або використовує природні умови), що є необхідними задля вияву певних властивостей, то спостереження розглядається як систематичне направлене сприйняття того або іншого об'єкта (чи явища) задля отримання первинної інформації без впливу на останній.

У науковій літературі отримання статистичних даних в такі способи визначають як результат спостереження (вимірювання, випробування, досліду чи аналізу).

Обробку результатів спостереження не завжди можливо виконати однозначно, тому що має місце:

- ✦ обмеженість числа виконаних спостережень, що не дозволяють з високою мірою достовірності оцінити форму й вид закону розподілу, а значить, й вибрати необхідний критерій оцінки результату на наявність неточності;
- ✦ наявність особливостей дослідження певного об'єкта (чи процесу), показники (параметри) якого утворюють випадкову сукупність (вибірку).

Виходячи з цього, *метою* даної роботи є визначення проблемних питань щодо обробки результатів статистичного спостереження, обов'язковість розгляду яких нехтується при аналізі складних соціально-економічних явищ чи процесів.

При обробці результатів спостереження однозначно мають місце як проблеми в описі даних та їх оцінюванні, так й проблеми перевірки гіпотез. Аналіз наукових джерел дозволив зробити висновок, що відповідно до сфери проведення спостереження співвідношення між блоками проблемних питань обробки результатів статистичного спостереження має суттєві коливання. Так, в економічній сфері суттєвим є відсоток проблем, пов'язаних з описом даних і перевіркою гіпотез; у соціальній – з їх оцінкою; в технічній, залежно від складності вирішення завдань, – має місце відносно пропорційний їх розподіл.

Не виходячи зі сфери проведення статистичного спостереження, процес збирання вихідних даних пов'язано з

їх вимірюванням, тобто визначенням дійсних значень постійної чи змінної величини, що вимірюється. Результатом як однократних, так й багатократних вимірів є реалізація випадкової величини. Вимір є закінченим, якщо повністю визначено не тільки значення величини, що вимірюється, але й можлива міра його відхилення від дійсної величини.

Під спостереженням при вимірі розуміють операції, що проводяться при вимірі та які мають за мету своєчасно й правильно провести відлік. Наприклад, якщо компанія використовує схему роботи «на склад», то на питання, що здійснюється раніше: закупівля товару чи її продаж, можуть бути отримано дві різні відповіді. Якщо товар є на складі, то його закупівля є первинною порівняно з продажем, а якщо товару на складі немає й клієнт згоден чекати, поки буде здійснено закупівлю цього товару, то процес продажу починається раніше в часі, ніж закупівля, а закінчується пізніше. Таким чином, результат спостереження містить всі види неточностей, що притаманні процедурі виміру.

Загалом, проблемою при обробці результатів спостереження є те, що неточності виміру частіше всього розглядається лише як зона невизначеності результатів. Так, лише виходячи з правил округлення розрахункового значення неточності й отриманого результату виміру, кількість значимих цифр у числовому значенні результату виміру дає змогу приблизно визначити його точність. Це пов'язано з тим, що гранична помилка, що зумовлена округленням, дорівнює половині одиниці посліднього розряду числового значення результату виміру. Оцінені межі похибки округлення характеризують вплив округлення на точність результату виміру. Крім того, ці дані дозволяють орієнтуватися в мінімально необхідному числі значимих цифр для запису результату виміру за умов його заданої точності.

При проведенні статистичного спостереження в будь-якій сфері особливою проблемою стає вибір способу визначення вибіркової характеристики та формування основи вибірки, тому що у відповідності до потреб одномірної статистики опис статистичних даних дається за допомогою частот, а у випадку багатовимірного статистичного аналізу, статистики випадкових процесів і часових рядів, статистики об'єктів нечислової природи, вихідні статистичні дані – це вибірка або кілька вибірок.

Як одна з основних характеристик виміру, неточність повинна бути обов'язково оцінена.

Основою для оцінювання невідомого параметра розподілу є певна оцінка. Якщо оцінювання – це дія, то оцінка – статистика (функція від результатів спостережень), що використовується в процесі зазначеної дії або є його результатом. Практично всі оцінки параметрів, які використовуються в імовірнісно-статистичних методах прийняття рішень, є або незміщеними, або асимптотично незміщеними.

Проблема оцінювання параметрів розподілу часто пов'язана з наявністю аномальних спостережень, які необхідно відбракувати. Реалізувати відбракування спостережень у загальному випадку виявляється зовсім не просто, тому що спостереження, аномальні з позицій одного закону розподілу, є природним проявом закономірностей другого. Якщо немає надійної процедури відбракування, вийти з положення можна:

- ✦ скорочуючи вибірку – відкидаючи певну частину мінімальних і/або максимальних спостережень, і по частині оцінюючи параметри розподілу;
- ✦ вінзуючи вибірку – присвоюючи однакові значення перед процедурою оцінювання всім спостереженнями лівіше і/або правіше певних значень;
- ✦ цензуючи вибірку – фіксуєчи лише факт попадання у відповідний інтервал, опускаючи конкретні значення спостережень, що потрапили лівіше і/або правіше певних значень.

Аналіз застосування зазначених процедур у різних сферах проведення статистичного спостереження показав, що дві перші процедури можуть бути використані, коли припускається формування нової генеральної сукупності, до якої належить видозмінена вибірка, а процедура цензурування має більш широкий спектр застосування. Але при наявності аномальних спостережень і відхилень передбачуваної моделі від закону, що реально спостерігається, застосування цензурування, одностороннього або двостороннього, рідко приносить бажаний ефект.

Виходячи з того, що одним з найбільш ефективних методів оцінювання параметрів розподілів є метод максимальної правдоподібності, розглянемо вплив аномальних помилок на ОМП параметрів при нормальному розподілу та проведемо оцінку за вибіркою, що належить до іншого розподілу (табл. 1).

жити розподілу Лапласа, що має більш «важкі» хвости, ніж нормальний, не можна говорити про близькість емпіричної та теоретичної функції розподілу.

Природно, що використання отриманих в цих двох ситуаціях нормальних законів як моделей спостережуваних вибірок не є коректним.

Різно знизити вплив аномальних спостережень, а іноді практично виключити наслідки присутності їх у вибірці, крім того, різко знизити вплив на оцінки параметрів і відхилень виду закону, що спостерігається від передбачуваного, дозволяє групування вибірки.

Результати оцінювання за згрупованими даними параметрів нормального закону за вибіркою, що належить розподілу Лапласа, показали, що спостерігається деяка близькість емпіричної функції розподілу та функції розподілу нормального закону.

Але при групуванні спостережень мають місце втрати інформації про параметри закону розподілу. Чим більше втрати, тим більша асимптотична дисперсія оцінки. Реально ж, за наявності у вибірці неодноразово згадуваних відхилень, дисперсія оцінок по згрупованим даним виявляється менше, ніж по не згрупованим. Тобто внесок у дисперсію від втрат у асимптоті виявляється незрівнянно малим у порівнянні з внеском, пов'язаним з наявністю відхилень.

Проблемним залишається питання здійснення самого угруповання. Досліднику при проведенні спостереження у будь-якій сфері необхідно пам'ятати, що згрупувати дані необхідно так, щоб у згрупованій вибірці залишилася інформація про закон і його параметри.

На практиці, на інтервали розбивають область, яка визначається розмахом вибірки. Це можуть бути: інтервали однакової довжини, або інтервали рівної частоти, або асимптотично оптимальні інтервали чи інтервали, сформовані за будь-яким іншим принципом.

Найбільш стійкими до відхилень виявляються оцінки при розбитті вибірки на інтервали рівної частоти

Таблиця 1

Результати оцінювання параметрів нормального розподілу та подальшого аналізу

| Результати статистичного аналізу | Обчислені значення ОМП | |
|---|------------------------|----------|
| | μ | σ |
| Початкова вибірка за не згрупованими даними | -0,0102 | 1,0189 |
| Вибірка з аномальними спостереженнями | -0,0203 | 1,1614 |
| Вибірка, що належить розподілу Лапласа | 0,0029 | 1,4281 |
| Вибірка за згрупованими даними з аномальними спостереженнями | -0,0061 | 1,0397 |
| Вибірка за згрупованими даними, що належить розподілу Лапласа | 0,0169 | 1,1947 |

Щоб знизити вплив аномальних помилок і відхилень вибірок, що спостерігаються від передбачуваного закону на оцінки обчислюваних параметрів, буде доречним використовувати перед обчисленням оцінок параметрів процедуру попереднього групування спостережень.

Отримана ОМП параметрів нормального розподілу свідчить, що особливо істотна наявність аномальних спостережень вплинула на оцінку середньоквадратичного відхилення.

Розглянувши ситуацію з використанням нормального розподілу в ситуації, коли насправді вибірка нале-

(ймовірності). У разі асимптотично оптимального групування, втрати інформації про параметри закону розподілу, пов'язані з групуванням, істотно менше, ніж при рівноймовірному. Якщо відомо, що відхилення від припущень у вибірці мінімальні, то використання таблиць асимптотично оптимального групування дозволяє різко скоротити обсяги даних, що спостерігалися, без істотної втрати інформації про закон розподілу. У загальному випадку – слід чекати більшої чутливості оцінок щодо відхилень від припущень.

З іншого боку, досить часто інформація, що пов'язана із засміченням вибірки, чинить менший вплив на оцінки, ніж втрати інформації від групування при асимптотично оптимальному групуванні. У деяких випадках оцінки з використанням асимптотично оптимального групування виявляються так само стійкими, як і при рівномірному. І при цьому показують кращі результати. Тому буде корисним обчислювати дві оцінки за згрупованими даними з використанням як оптимального, так і рівномірного групування, і зупинитися необхідно на тій оцінці, котра дає краще узгодження з вихідною вибіркою.

Найбільш реальною з позицій практичної діяльності є ситуація, коли у вибірці немає аномальних спостережень, але закон розподілу, що спостерігається, відрізняється від передбачуваного. Реальність даної ситуації визначена тим, що безліч законів розподілу ймовірностей нескінченна, а кількість моделей, що використовуються на практиці для опису випадкових величин, що спостерігаються дуже обмежена.

Таким чином, чим істотніше від передбачуваної моделі відрізняється від реально спостережуваного закону, тим сильніше це відбивається на оцінках параметрів.

В аналітичних розрахунках буде доцільним використання оцінки за згрупованими даними, разом з оцінками за незгрупованими спостереженнями. Їх якість залежить від ступеня засміченості вибірки аномальними спостереженнями чи близькості до передбачуваного законом розподілу.

При статистичній обробці результатів спостережень: аналізі даних, обробці результатів вимірювань, дослідженні надійності та контролю якості основну увагу повинно бути приділено способам розбиття вибірки на інтервали, що забезпечує максимальну потужність критерію. ■

ЛІТЕРАТУРА

1. Левин, Давид М., Стефан, Давид, Кребиль, Тимоти С., Беренсон, Марк Л. Статистика для менеджеров с использованием Microsoft Excel. – 4-е изд. : Пер. с англ. – М. : Издательский дом «Вильямс», 2004. – 1312 с.

2. Лемешко Б. Ю. О распределениях статистик непараметрических критериев согласия при оценивании по выборкам параметров наблюдаемых законов / Б. Ю. Лемешко, С. Н. Постовалов // Заводская лаборатория. – 1998. – Т. 64. – № 3. – С. 61 – 72.

3. Lemeshko B. Y., Lemeshko S. B. Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II, Measurement Techniques, 2009. – № 52:8. – С. 799 – 812.

4. Новицкий П. В. Оценка погрешностей результатов измерений / П. В. Новицкий, И. А. Зограф. – Л. : Энергоатомиздат, 1985. – 247 с.

5. Орлов А. И. Неустойчивость параметрических методов отбраковки резко выделяющихся наблюдений / А. И. Орлов // Заводская лаборатория. – 1992. – Т. 58. – № 7. – С. 40 – 42.

6. Третьяк Л. Н. Обработка прямых измерений с многократными наблюдениями : Учебное пособие / Л. Н. Третьяк. – Оренбург : ИПК ОГУ, 2002. – 60 с.

7. Chandra M., Singpurwalla N. D., Stephens M. A. Statistics for Test of Fit for the Extrem-Value and Weibull Distribution // J. Am. Statist. Assoc. – 1981. – V. 76. – P. 375.