УДК 519.25

В.Є. Бахрушин, І.О. Дудко

УТОЧНЕННЯ МОДЕЛЕЙ НОРМАЛЬНОГО РОЗПОДІЛУ НА ОСНОВІ МІНІМІЗАЦІЇ КРИТЕРІЮ КОЛМОГОРОВА-СМИРНОВА

Аннотация. В статье рассмотрен вопрос об оценивании критических значений критерия согласия Колмогорова-Смирнова для случая, когда параметры модели нормального распределения рассчитываются путем минимизации расчетного значения этого критерия.

Вступ. Традиційна процедура ідентифікації моделей розподілу складається з трьох етапів [1, 2]: формування гіпотези про ймовірнісний закон розподілу; оцінювання параметрів обраної моделі; перевірка адекватності моделі за допомогою статистичних критеріїв. Останній етап у [3] рекомендовано виконувати з використанням критеріїв типу омега-квадрат та Колмогорова-Смирнова. Втім існує проблема вибору методів оцінювання параметрів моделі. Зазвичай при використанні моделей нормального розподілу для цього беруть вибіркові середнє арифметичне та стандартне відхилення. У [2] показано, що модель може бути істотно покращена, якщо здійснити мінімізацію розрахункового значення критерію Колмогорова-Смирнова, використовуючи ці параметри як початкове наближення. Але при цьому постає питання про те, з якими критичними значеннями слід порівнювати отриману величину для того, щоб прийняти або відхилити нульову гіпотезу про відповідність даних отриманій моделі. Можна очкувати, що критичне значення залежатиме у цьому випадку від вибору конкретного методу мінімізації, а також від того, яку саме гіпотезу перевіряють – про відповідність даних нормальному розподілу (незалежно від параметрів моделі), чи про їх відповідність моделі з визначеними на попередньому етапі параметрами.

Одним з шляхів вирішення цієї проблеми може бути встановлення статистичного зв'язку між розрахунковими значеннями, одержуваними при традиційній методиці ідентифікації, та значеннями,

які є результатами мінімізації критерію Колмогорова-Смирнова для тих самих вибірок.

Постановка завдань. Метою роботи було визначення характеристик статистичного зв'язку між розрахунковими значеннями критерію Колмогорова-Смирнова, одержуваних при використанні різних методик ідентифікації моделі розподілу. Для цього було взято дві серії вибірок, одна з яких була згенерована за допомогою генератора випадкових чисел, а друга – являла собою реальні емпіричні дані. Потім досліджено кореляцію між розрахунковими значеннями критерію, отриманих за різними методиками для обох серій вибірок, й побудовано відповідні регресійні моделі.

Критерій згоди Колмогорова-Смирнова. На практиці постає питання, чи погоджуються результати спостережень з функцією розподілу сімейства $\{F(x;\Theta), \theta \in \Theta\}$ при визначеному $\theta = \theta_0$. Відповідні гіпотези називають гіпотезами згоди, а методи їх перевірки – критеріями згоди [3]. Умови застосування непараметричних критеріїв згоди при перевірці простих і складних гіпотез про вигляд функції розподілу зустрічається в роботах [1–6].

Одним з найбільш популярних непараметричних критеріїв згоди є критерій Колмогорова-Смирнова, розроблений А.М. Колмогоровим та М.В. Смирновим в 1930-х роках. Його сутність полягає у визначенні максимального за модулем відхилення моделі розподілу від наявних емпіричних точок та його порівнянні з критичним значенням [1, 7]. Нехай $F_n(x)$ – емпірична функція розподілу випадкової величини x, що подана у вигляді вибірки $x_1 \le x_2 \ldots \le x_n$:

$$F_n(x) = \begin{cases} 0, & x < x_1; \\ \dfrac{1}{n}, & x_i \le x \le x_{i+1},\ 1 \le i \le n-1; \\ 1, & x \ge x_n. \end{cases}$$

Далі, залежно від наявної інформації про функцію розподілу, вирізняють два шляхи.

1) Якщо відомі значення параметрів $\theta = \theta_0$ і функція розподілу $F(x;\Theta)$ неперервна, то для перевірки двосторонньої нульової гіпотези

 ISSN 1562-9945

$H_0 : F_n(x) = F(x)$, розраховують максимальну відстань між емпіричною і теоретичною функціями розподілу:

$$D_n = \sup_{|x|<\infty} \left| F_n(x) - F(x, \theta_0) \right|,$$

$$D_n^+ = \sup_{|x|<\infty} \left(F_n(x) - F(x, \theta_0) \right),$$

$$D_n^- = -\inf_{|x|<\infty} \left(F_n(x) - F(x, \theta_0) \right).$$

На практиці для цього використовують формули:

$$D_n^+ = \max_{1 \le i \le n} \left(\frac{i}{n} - F(x_i, \theta_0) \right),$$

$$D_n^- = \max_{1 \le i \le n} \left(F(x, \theta_0) - \frac{i-1}{n} \right),$$

$$D_n = \max_{1 \le i \le n} \left(D_n^+, D_n^- \right).$$

У випадку односторонньої нульової гіпотези беруть $D_n = D_n^+$.

Емпіричне значення критерію Колмогорова-Смирнова визначають за формулою $\lambda = \sqrt{n} D_n$ та порівнюють з критичним значенням для заданого рівня значущості.

2) Якщо теоретичне значення параметра θ_0 невідоме, то використовують статистику

$$\lambda^* = \sqrt{n} D_n \left(\theta^* \right),$$

де на відміну від істинних значень параметрів θ_0 використовують вектор їх вибіркових оцінок θ^*. Визначення величини $D_n \left(\theta^* \right)$ здійснюють так саме, як і у першому випадку.

Нульову гіпотезу про відповідність вибірки заданій моделі розподілу приймають, якщо розрахункове значення критерію є меншим, ніж критичне для заданого рівня значущості. В іншому випадку цю гіпотезу відхиляють.

Критичні значення статистики Колмогорова-Смирнова, визначають залежно від способу визначення параметрів моделі (табл. 1) [3].

Таблиця 1

Критичні значення статистики Колмогорова-Смирнова

Довірчий рівень, p	0,85	0,90	0,95	0,975	0,99
Критичне значення для D_n	1,138	1,224	1,358	1,480	1,626
Критичне значення для $D_n\left(\theta^*\right)$	0,775	0,819	0,895	0,955	1,035

Зважаючи на те, що критичні значення критерію Колмогорова-Смирнова не залежать від типу невідомого розподілу, а також те, що як відстань між $F_n(x)$ та $F(x)$ беруть максимальне відхилення, можна використовувати процедуру перевірки гіпотези згоди при пошуку довірчих меж неперервної функції розподілу [4]. Незалежно від обраної моделі (теоретичної функції) розподілу:

$$P\left\{F_n(x) - d_\alpha \le F(x) \le F_n(x) + d_\alpha\right\} = 1 - \alpha, \text{ для всіх } x, \qquad (1)$$

де d_α – критичне значення D_n для рівня значущості α.

Формула (1) означає, що теоретична функція $F(x)$ цілком знаходиться всередині інтервалу $\pm d_\alpha$, утвореного навколо обраної точки $F_n(x)$.

Можна отримати оцінки обсягу вибірки, потрібного для апроксимації функції розподілу з необхідною точністю. При $\alpha \le 0,2$ та $n \ge 80$:

$$d_\alpha \approx \sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}} \Big/ \sqrt{n}.$$

При збільшенні обсягу вибірки n, розподіл статистик D_n^+, D_n^-, D_n наближується до граничних. Для невеликих обсягів вибірок різниця між граничним і вибірковим розподілом може бути суттєвою. Вирішення цієї проблеми полягає у використанні нових статистик $\tilde{D}_n^+$, $\tilde{D}_n^-$, $\tilde{D}_n$, розподіли яких краще узгоджуються з граничними за малих n. Формули для розрахунку цих статистик представлені в таблиці 2 [4].

 ISSN 1562-9945

Таблиця 2

Формули модифікованих статистик

Вихідна статистика	Модифікована статистика
	Верхній хвіст розподілу
$\sqrt{n}D_n^+$	$\tilde{D}_n^+ = D_n^+\left(\sqrt{n} + 0,12 + 0,11/\sqrt{n}\right)$
$\sqrt{n}D_n^-$	$\tilde{D}_n^- = D_n^-\left(\sqrt{n} + 0,12 + 0,11/\sqrt{n}\right)$
$\sqrt{n}D_n$	$\tilde{D}_n = D_n\left(\sqrt{n} + 0,12 + 0,11/\sqrt{n}\right)$
	Нижній хвіст розподілу
$\sqrt{n}D_n$	$\tilde{D}_n = D_n\left(\sqrt{n} + 0,275 + 0,04\sqrt{n}\right)$

Однією з властивостей критерію Колмогорова-Смирнова є його стійкість проти будь-якої альтернативної гіпотези, тобто будь-яка відмінність розподілу вибірки від теоретичного закону буде виявлена, за умови великої кількості спостережень.

Процедура мінімізації розрахункових значень. При ідентифікації моделей розподілу важливим завданням є підбір такої моделі з заданого класу, яка б найкраще відповідала наявним емпіричним даним. Для вирішення цієї проблеми одним з авторів в роботах [1, 2] було запропоновано використовувати процедуру мінімізації розрахункових значень статистичних критеріїв із застосуванням параметрів моделі як змінних, які можна варіювати під час цієї процедури. При цьому як початкове наближення використовуються вибіркові оцінки параметрів, зокрема для моделі нормального розподілу – вибіркові середнє арифметичне й стандартне відхилення. При використанні критерію Колмогорова-Смирнова, це означає, що необхідно підібрати такі параметри розподілу, щоб значення λ^* було найменшим.

Реалізацію запропонованої процедури можна здійснювати за допомогою будь-яких програмних засобів, де реалізовано сучасні алгоритми багатовимірної нелінійної оптимізації. Зокрема при застосуванні табличного процесора Microsoft Excel для цього можна використовувати надбудову «Пошук розв'язку» [8]. У цьому випадку як цільову необхідно взяти ту комірку, в якій міститься значення λ^*, вибрати пошук мінімального значення та змінювати комірки, що містять

значення параметрів моделі. Діалогове вікно цієї процедури зображено на рисунку 1.

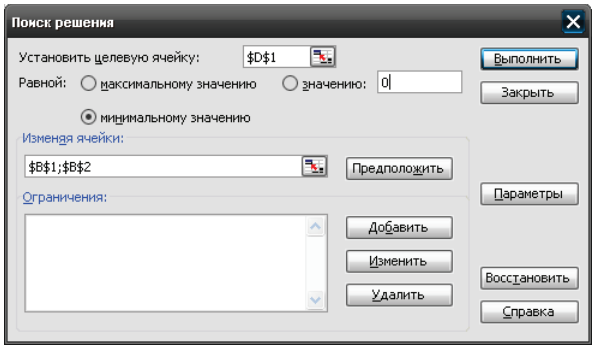

Рисунок 1 - Діалогове вікно «Пошук розв'язку»

Також є можливість встановити додаткові параметри пошуку розв'язку в діалоговому вікні, що викликається кнопкою «параметри».

Процедура мінімізації дає змогу суттєво знизити розрахункові значення критерію, тобто покращити одержану модель, але постає проблема того, які критичні значення слід використовувати для прийняття чи відхилення гіпотези про її відповідність розподілу наявним емпіричним даним.

Результати дослідження. Було досліджено дві серії вибірок. Перша серія являла собою 60 вибірок обсягом по 1000 елементів кожна, отриманих за допомогою вбудованого генератора випадкових чисел MS Excel за нормальним розподілом із середнім значенням $\bar{x} = 0$ та стандартними відхиленнями σ від 1 до 10. Після цього для кожної вибірки визначали вибіркові середнє арифметичне й стандартне відхилення та розраховували значення критерію Колмогорова-Смирнова λ^*. Далі виконували описану вище процедуру мінімізації розрахункового значення й одержували його нове значення $\lambda^*_{опт.}$.

Як приклад, наведемо результати, отримані для однієї з вибірок, згенерованої за нормальним законом розподілу з параметрами $n = 1000$, $\bar{x}_0 = 0$, $\sigma = 4$. Відповідні вибіркові значення дорівнювали: $\bar{x} = 0,115$, $s = 3,915$, $x_{min} = -13,339$, $x_{max} = 12,353$, а результати розрахунку критерію Колмогорова-Смирнова були такими: $D_n = 0,038$, $\lambda = 1,204$, $D_n^* = 0,024$, $\lambda^* = 0,760$.

 ISSN 1562-9945

Після застосування процедурі «Пошук розв'язку» знайдено нові значення: $\bar{x} = 0,125$, $s = 3,801$, $D_n^* = 0,017$, $\lambda_{опт.}^* = 0,548$. Отримане $\lambda_{опт.}^*$ є значно меншим ніж λ^*.

Далі розглянемо результати по всім вибіркам. На рис. 2. показано зв'язок між λ^* та $\lambda_{опт.}^*$ для вибірок, що були згенеровані в MS Excel.

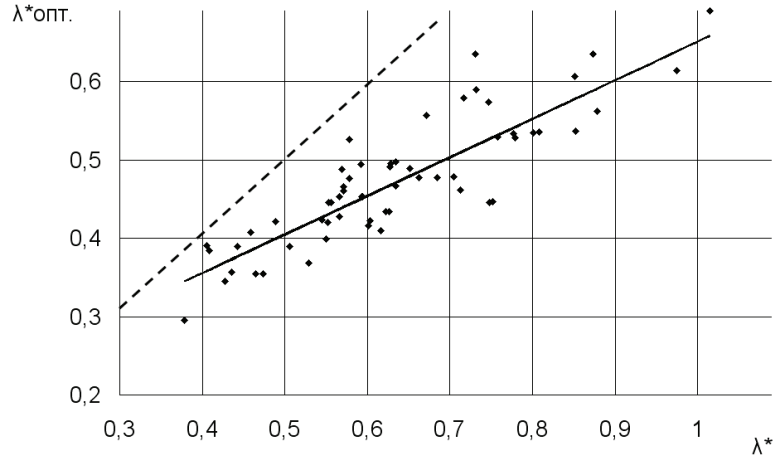

Рисунок2 - Зв'язок між λ^* та $\lambda_{опт.}^*$ для вибірок, згенерованих в MS Excel

Коефіцієнт кореляції Пірсона між λ^* та $\lambda_{опт.}^*$ дорівнює 0,86, що свідчить про наявність сильного лінійного зв'язку між досліджуваними змінними. При перевірці значущості коефіцієнта кореляції r для $\alpha = 0,05$ та кількості степенів вільності $k = 58$, отримали розрахункове значення t – критерію Стьюдента $t = 13,01$, критичне значення $t_{0,95;\,58} = 2,00$. З $t > t_{0,95;\,58}$ випливає, що коефіцієнт кореляції між λ^* та $\lambda_{опт.}^*$ значимо відрізняється від нуля. Значення коефіцієнта кореляції ρ з надійністю 0,95 міститься в інтервалі $0,780 \le \rho \le 0,916$. Рівняння регресії має вигляд: $\lambda_{опт.\lambda^*}^* = 0,492\lambda^* + 0,159$.

Друга серія вибірок являла собою дані про показники складності та коефіцієнти дискримінації тестових завдань ЗНО 2009 – 2011, а також підсумкові й окремі показники низки міжнародних університетських рейтингів. Загалом ця серія містила 118 вибірок обсягом від 32 до 400 елементів. Процедура аналізу була такою самою, як і у попередньому випадку.

На рис. 3. показано зв'язок між λ^* та $\lambda^*_{\text{опт.}}$ для цієї серії.

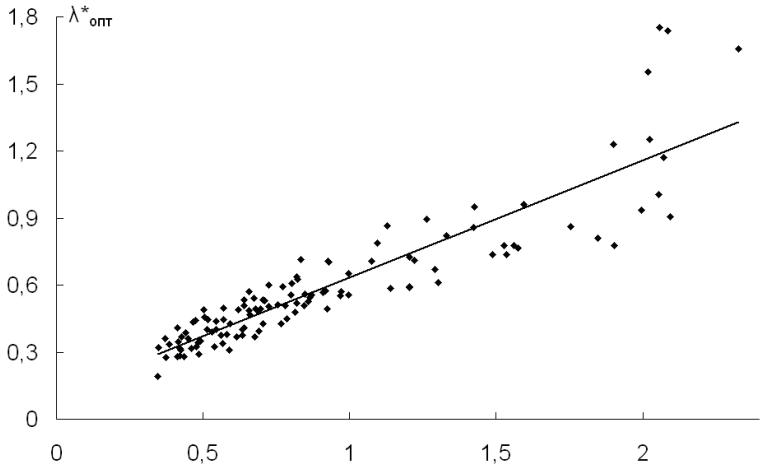

Рисунок 3 - Зв'язок між λ^* та $\lambda^*_{\text{опт.}}$ для серії вибірок, що містять освітні показники

Коефіцієнт кореляції Пірсона між λ^* та $\lambda^*_{\text{опт.}}$ у цьому випадку є трохи вищим і дорівнює $0,90$, що також відповідає гіпотезі про наявність сильного лінійного зв'язку між досліджуваними показниками. При перевірці значущості коефіцієнта кореляції r для $\alpha = 0,05$ та кількості степенів вільності $k = 116$, отримали розрахункове значення $t-$критерію Стьюдента $t = 13,01$, критичне значення $t_{0,95;\,58} = 2,00$. З $t > t_{0,95;\,58}$, як і у попередньому випадку, випливає, що коефіцієнт кореляції між λ^* та $\lambda^*_{\text{опт.}}$ значимо відрізняється від нуля. Значення коефіцієнта кореляції ρ з надійністю $0,95$ міститься в інтервалі $0,780 \le \rho \le 0,916$. Рівняння регресії має вигляд: $\lambda^*_{\text{опт.}_{\lambda^*}} = 0,525\lambda^* + 0,109$. При цьому довірчі рівні коефіцієнтів моделі при рівні значущості $0,05$ істотно перетинаються. Для коефіцієнта при λ^* для першої моделі маємо довірчий інтервал $a \in [0,416;\, 0,568]$, а для другої – $a \in [0,477;\, 0,572]$. Відповідно, значення вільного члена для першої моделі перебуває в діапазоні $b \in [0,109;\, 0,208]$, а для другої – $a \in [0,059;\, 0,157]$.

Отримані залежності дають змогу приблизно встановити критичні значення критерію Колмогорова-Смирнова у випадку, коли для більш точного визначення параметрів моделі нормального розподілу

використовується мінімізація розрахункового значення цього критерію. Так, рівню значущості 0,05 у цьому випадку буде відповідати критичне значення приблизно 0,57 – 0,6, а рівню значущості 0,1 – критичне значення 0,53 – 0,56.

Висновки. Показано можливість оцінювання критичних значень критерію Колмогорова-Смирнова у випадку, коли значення параметрів моделі нормального розподілу уточнюють шляхом мінімізації розрахункового значення цього критерію. Отримано статистичну модель зв'язку між цими значеннями й відповідними розрахунковими значеннями, визначеними для випадку, коли використовують вибіркові оцінки цих параметрів. У подальшому планується уточнити параметри зв'язку на вибірках великого обсягу, а також перевірити аналогічні зв'язки при інших моделях розподілу.

ЛІТЕРАТУРА

1. Бахрушин В.Є. Методи аналізу даних / В.Є. Бахрушин – Запоріжжя: КПУ, 2011. – 268 с.

2. Бахрушин В.Е. Проблемы идентификации моделей распределения случайных величин с применением современного программного обеспечения / В.Є. Бахрушин // Успехи современного естествознания. – 2011. – № 11. – С. 50 – 54.

3. Орлов А.И. Прикладная статистика. Учебник / А.И. Орлов. – М.: Издательство «Экзамен», 2004. – 656 с.

4. Айвазян А.С. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1983. – 471 с.

5. Лемешко Б.Ю. О задаче идентификации закона распределения случайной составляющей погрешности измерений / Б.Ю. Лемешко // Метрология. – 2004. – № 7. – С. 8 – 18.

6. Лемешко Б.Ю. О правилах проверки согласия опытного распределения с теоретическим / Б.Ю. Лемешко, С.Н. Постовалов // Методы менеджмента качества. Надежность и контроль качества. – 1999. – № 11. – С. 34 – 43.

7. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных сотрудников / А.И. Кобзарь. – М.: ФИЗМАТЛИТ, 2006. – 816 с.

8. Лялин В.С. Статистика: теория и практика в Excel: учеб. пособие / В.С. Лялин, И.Г. Зверева, Н.Г. Никифорова. – М.: Финансы и статистика; ИНФРА–М, 2012. – 448 с.: ил.