

Олександр Григорович ХАНІН

кандидат фізико-математичних наук,
доцент кафедри математичного аналізу,
Східноєвропейський національний університет імені Лесі Українки
E-mail: a.hanin@bigmir.net

МЕТОДОЛОГІЧНІ ОСОБЛИВОСТІ ЗАСТОСУВАННЯ КРИТЕРІЮ УЗГОДЖЕНОСТІ χ^2 В ПРАКТИЧНИХ ЗАДАЧАХ СОЦІОЛОГІЇ, ЕКОНОМІКИ ТА МАРКЕТИНГУ

Ханін, О. Г. Методологічні особливості застосування критерію узгодженості χ^2 в практичних задачах соціології, економіки і маркетингу [Текст] / Олександр Григорович Ханін // Економічний аналіз : зб. наук. праць / Тернопільський національний економічний університет; редкол. : В. А. Дерій (голов. ред.) та ін. – Тернопіль : Видавничо-поліграфічний центр Тернопільського національного економічного університету “Економічна думка”, 2015. – Том 22. – № 1. – С. 67-70. – ISSN 1993-0259.

Анотація

У багатьох практичних задачах соціології, економіки та маркетингу виникає потреба порівняння ймовірнісних розподілів різних якісних факторів. Це можуть бути смакові переваги користувачів, джерела пошуку інформації щодо товару чи послуги, структура збуту продукції, фінансово-економічні результати діяльності підприємств у різноманітних розрізах тощо. Розподіли можуть залежати від різних соціальних та фінансово-економічних чинників: віку, освіти, рівня доходів, належності до певних професійних груп, місцевих особливостей і т. д.

У статті показано, що класичний метод перевірки статистичної гіпотези про відмінність між собою двох вибірових розподілів факторів за допомогою критерію узгодженості χ^2 , створеного для порівняння емпіричного та теоретичного розподілів, не є повною мірою коректним, тому запропоновано і розглянуто на конкретному прикладі методику використання критерію χ^2 для практичного розв'язанні подібних задач.

Ключові слова: теоретичний та вибіровий розподіл; біноміальний розподіл; розподіл Стьюдента; критерій узгодженості χ^2 ; довірчий інтервал; інструмент Excel «Пошук розв'язків».

Oleksandr Gryhorovych KHANIN

PhD in Physical and Mathematical Sciences,
Associate Professor,
Department of Mathematical Analysis,
Lesya Ukrainka Eastern European National University
E-mail: a.hanin@bigmir.net

METHODOLOGICAL FEATURES OF χ^2 CONSISTENCY CRITERIA USING IN THE PRACTICAL PROBLEMS OF SOCIOLOGY, ECONOMICS AND MARKETING

Abstract

There is a need to compare probability of distributions of various qualitative factors in many practical problems of sociology, economics and marketing. They may be the taste preferences of customers, the sources of the searching for information about the goods or services, sales structure, financial and economic results in different sections etc. The distribution may depend on different social, financial and economic factors: age, education, income, belonging to certain professional groups, local features, etc.

The article shows that the classic method of testing of statistical hypotheses about the difference between sample distributions of some factors using χ^2 consistency criterion, created for comparison of empirical and theoretical distributions, is not fully correct. Therefore, it is proposed the method of using χ^2 test for the practical solution of similar problems.

Keywords: theoretical and sample distribution; binomial distribution; Student distribution; consistency criterion χ^2 ; confidence intervals; Excel tool «Search for solutions».

Критерій χ^2 був розроблений К. Пірсоном [1] з метою порівняння емпіричного розподілу з відомим теоретичним. Р. Фішер [2; 3] удосконалив цей критерій на випадок, коли вид теоретичного розподілу відомий, але розподіл залежить від кількох невідомих параметрів. Якщо емпіричний розподіл r факторів наближається до відомого теоретичного із зростанням об'єму вибірки, то введена Пірсоном відстань між емпіричними частотами та відповідними теоретичними ймовірностями має при достатньо великому об'ємі вибірки n розподіл, близький до розподілу χ^2 з $r-1$ ступенями вільності. Але на практиці ми зустрічаємося з дещо іншою ситуацією: вид теоретичного розподілу нам невідомий і необхідно порівняти між собою не емпіричний та теоретичний, а два емпіричних розподіли. Припустимо, обидва емпіричних розподіли r факторів побудовані за вибірками об'ємів n_1 і n_2 , відповідно, та коли n_1 і n_2 необмежено зростають, наближаються до одного і того самого невідомого нам теоретичного розподілу. Але, якщо визначити між ними відстань Пірсона, її розподіл залишиться нам невідомим, зокрема, ми не зможемо гарантувати його близькість до розподілу χ^2 . Тому виникає необхідність дещо модифікувати критерій χ^2 для застосування у нашому випадку.

Мета статті – модифікувати критерій Пірсона так, щоб його можна було використати для порівняння двох емпіричних розподілів, не збільшуючи помилку 1-го роду, тобто ймовірність помилкового визнання двох емпіричних розподілів різними.

Нехай досліджується пріоритетність факторів, що впливають на вибір певного приватного медичного закладу, залежно від вікової групи, до якої належить потенційний пацієнт. Розглянемо для прикладу дві вікові групи: до 30 років та 30-50 років.

Таблиця 1. Розподіл частот факторів вибору приватного медичного закладу по вікових групах

Фактори	До 30 років	30-50 років
	К-ть осіб	К-ть осіб
Ціна	34	190
Імідж лікаря	103	174
Рекомендації знайомих	24	73
Репутація закладу	90	126
Реклама	14	13
Направлення дільничного лікаря	18	105
Усього	283	681

Виникає задача прийняття обґрунтованого рішення про однакову чи різну структуру пріоритетності факторів вибору приватного медичного закладу залежно від віку потенційного пацієнта. Очевидним результатом такого аналізу буде вироблення однакової або різних маркетингових стратегій.

Зрозуміло, що аналогічні задачі виникають при порівняльному аналізі структури попиту на різні групи продукції залежно від соціальних, географічних або інших особливостей потенційних клієнтів або при порівнянні відповідної структури збуту. Спроба розв'язати таку задачу за допомогою критерію χ^2 не зовсім коректна, оскільки жоден з розподілів, з якими ми маємо справу, не є тим самим відомим теоретичним розподілом, до якого наближається емпіричний із зростанням об'єму вибірки, і відповідно відстань Пірсона між двома різними емпіричними розподілами не зобов'язана мати, як граничний, розподіл χ^2 .

Один з можливих шляхів виходу з ситуації може бути наступним. Розглянемо вікову групу «30-50 років» як «еталонну». Розподіл еталонної групи буде, в певному розумінні, аналогом теоретичного розподілу. Розіб'ємо всі фактори вибору приватного медичного закладу на дві групи, наприклад ціна та інші. У такому випадку вибіркового розподілу частот буде мати наступний вигляд.

Таблиця 2. Розподіл частот факторів вибору приватного медичного закладу по вікових групах у дихотомічній шкалі

Фактори	До 30 років	30-50 років
	К-ть осіб	К-ть осіб
Ціна	34	190
Інші	249	491
Всього	283	681

Умовно розглядатиме вибір фактору «Ціна» як «успіх», а іншого – як «невдача». У цьому випадку ми будемо мати вибірки з біноміального теоретичного розподілу з невідомою ймовірністю успіху p . Однак

для цієї ймовірності можна побудувати довірчий інтервал певної, наперед заданої надійності. За «еталонним» розподілом побудуємо асимптотичний двосторонній довірчий інтервал невідомої ймовірності «успіху» надійності 0,95. Для цього знайдемо частоту «успіху» $v_{усп}=190/681 \approx 0,279$, критичне значення двостороннього розподілу Стюдента з $n-1=680$ степенями вільності $t_{0.05} \approx 1,963$, вибіркочну оцінку невідомого стандартного відхилення ймовірності «успіху».

$$S_n = \sqrt{\frac{v_{усп}(1-v_{усп})}{n}} \approx 0,017.$$

Тоді [4, с. 400-401] права межа довірчого інтервалу дорівнює

$$v_{усп} + t_{0.05} \cdot S_n \approx 0,313,$$

ліва дорівнює

$$v_{усп} - t_{0.05} \cdot S_n \approx 0,245.$$

Тобто з ймовірністю 0,95 невідома ймовірність «успіху» $p \in (0,245; 0,313)$.

Повернемося тепер до даних еталонної групи («30-50 років») таблиці 1. Послідовно вважаючи за «успіх» кожну наступну категорію (імідж лікаря, рекомендації знайомих, репутація закладу, реклама, направлення дільничного лікаря), а решту об'єднавши поняттям «невдача», побудуємо, як ми це зробили вище, для відповідних ймовірностей «успіху» асимптотичні довірчі інтервали надійності 0,95.

Таблиця 3. Межі довірчих інтервалів для теоретичних ймовірностей, що відповідають кожному фактору вибору приватного медичного закладу

№ групи	Фактори	Ліва межа	Права межа
1	Ціна	0,245	0,313
2	Імідж/ім'я лікаря	0,223	0,288
3	Рекомендації знайомих	0,084	0,130
4	Репутація закладу	0,156	0,214
5	Реклама	0,009	0,029
6	Направлення лікарів	0,127	0,181

Як відомо [5, с. 454],

$$\chi^2 = \frac{1}{n} \sum_{i=1}^r \frac{(v_i - np_i)^2}{p_i}, \quad (1)$$

де r – кількість груп, на які розбито дані (кількість факторів, що розглядаються; в нашому випадку $r=6$),

n – об'єм вибірки (в нашому випадку $n=283$, тобто кількість респондентів віком до 30 років),

v_i – вибіркочна частота кожної групи (в нашому випадку – частоти, які містяться у стовпчику «До 30 років» таблиці 1),

p_i – теоретичні ймовірності, що навімання обраний респондент з генеральної сукупності з еталонної вікової групи вибере в якості найважливішого той чи інший фактор. У нашому випадку ці ймовірності невідомі, але знайдено довірчі інтервали для них (таблиця 3).

Оскільки нам невідомі теоретичні значення ймовірностей p_i , замінимо їх на такі значення з довірчих відрізків (тобто з довірчих інтервалів, разом із їх кінцями), які зроблять значення χ^2 у виразі (1) найменшим з можливих, тим самим при перевірці гіпотези про узгодженість розподілів ми мінімізуємо помилку 1-го роду.

У цьому полягає ідея методу. Залишається запропонувати шлях пошуку точок мінімуму функції (1) на замкненій області, яка становить декартовий добуток довірчих відрізків. Зауважимо, що довірчі відрізки не повинні містити нульові значення. Якщо це трапилось, тобто значення частот для деяких категорій (факторів) дуже малі, варто об'єднати ці категорії між собою або з іншими.

Знайти мінімум виразу (1) можна, використавши, наприклад, такий інструмент Excel, як «Пошук розв'язків». Відповідний аркуш Excel виглядатиме наступним чином (таблиця 4).

Цільова функція у таблиці 4 представляє собою вираз (1), а умова нормування вимагає рівності 1 суми шуканих ймовірностей. Необхідно також накласти умову позитивності шуканих ймовірностей.

Результатом роботи інструменту «Пошук розв'язків» є одночасно шукані ймовірності p_i , які мінімізують (1), а також саме мінімальне значення виразу (1), розподіленого у разі узгодженості

вибіркового розподілу (v_i/n , $i=1, \dots, b$) та теоретичного (із знайденими ймовірностями p_i , $i=1, \dots, b$) за законом χ^2 з $r-1=5$ ступенями вільності. Оскільки критичне значення такого розподілу на рівні значущості 0,05 становить $\approx 11,070$, а наше фактичне значення суттєво його перевищує ($\approx 281,004$), це свідчить про те, що розподіли частот факторів вибору приватного медичного закладу для вікових груп «До 30 років» та «30-50 років» є значуще різними, тобто респонденти цих вікових груп користуються різними пріоритетами при виборі приватного медичного закладу.

Таблиця 4. Результати роботи інструменту «Пошук розв'язків»

		Довірчі відрізки для невідомих теоретичних ймовірностей			
Фактори	Вибіркові частоти	Ліва межа	Права межа	Шукані ймовірності p	Нормування
Ціна	0,120	0,245	0,313	0,245	1,0000005
Імідж/ім'я лікаря	0,364	0,223	0,288	0,288	Об'єм вибірки n
Рекомендації знайомих	0,085	0,084	0,130	0,096	
Репутація закладу	0,318	0,156	0,214	0,214	283
Реклама	0,049	0,009	0,029	0,029	Цільова функція
Направлення лікарів	0,064	0,127	0,181	0,127	281,00432

Висновки та перспективи подальших розвідок

Запропонована методологія дає можливість використовувати критерій узгодженості χ^2 в багатьох практичних задачах у різних галузях соціології, психології, економіки, маркетингу тощо, коли виникає потреба у прийнятті на певному рівні значущості обґрунтованого рішення про узгодженість між собою тих чи інших вибіркових розподілів.

Список літератури

1. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling [Text] / K. Pearson // Philosophical Magazine. – 1900. – v. 50. – p. 157.
2. Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of P [Text] / R. A. Fisher // Journal of the Royal Statistical Society, 1922. – v. 85. – p. 87.
3. Fisher, R. A. The conditions under which χ^2 measures the discrepancy between observation and hypothesis. / R. A. Fisher // Journal of the Royal Statistical Society, 1924. – v. 87. – p. 442.
4. Сигел, Э. Практическая бизнес-статистика [Текст] / Э. Сигел. – М. : Вильямс, 2002. – С. 400-401.
5. Крамер, Г. Математические методы статистики [Текст] / Г. Крамер. – М. : Мир, 1976. – С. 454.

References

1. Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50, 157.
2. Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society, 85, 87.
3. Fisher, R. A. (1924). The conditions under which χ^2 measures the discrepancy between observation and hypothesis. Journal of the Royal Statistical Society, 87, 442.
4. Siegel, A. (2002). Practical Business Statistics. Moscow: Wilyams.
5. Kramer, H. (1976) Mathematical Methods of Statistics. Moscow: Mir.

Стаття надійшла до редакції 24.11.2015 р.