РОЗДІЛ «ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ»

УДК 519.233.3

АВРАМЕНКО В.І., к.т.н., доцент

Дніпродзержинський державний технічний університет

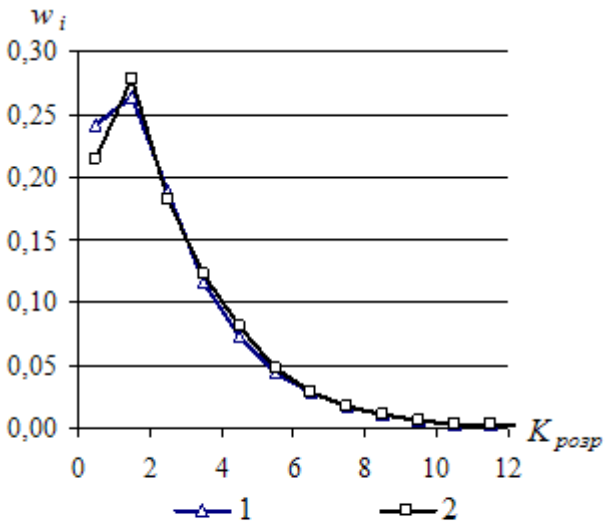
СТАТИСТИЧНЕ ДОСЛІДЖЕННЯ ВЛАСТИВОСТЕЙ КРИТЕРІЮ ЗГОДИ ПІРСОНА

Вступ. При статистичному аналізі масових явищ важливе значення має обґрунтований вибір закону розподілу сукупностей, що дозволяє правильно вибрати управлінське рішення. Для розв'язування задачі в межах статистичної перевірки статистичних гіпотез запропоновано кілька формалізованих критеріїв згоди. Одним з найбільш вживаних серед них є критерій Пірсона, який обчислюється за попередньо згрупованими даними. Розрахункове значення критерію $K_{розр}$ обчислюється за формулою

$$K_{\text{îçä}} = \sum_{i=1}^{k} \frac{(m_i - n_i)^2}{m_i},$$

де m_i і n_i – відповідно теоретичні і емпіричні частоти попадання до i-того інтервалу групування, k – кількість інтервалів групування, $\sum_{\text{³}=1}^{k} \grave{o}_{\text{³}} = \sum_{i=1}^{k} n_i = n$ – кількість спостережень у вибірці [1]. Дослідження властивостей критерію згоди Пірсона виконано методами статистичного моделювання з використанням стандартних функцій генерування випадкових величин СЛЧИСЛ, *rand* та *randG*.

Постановка задачі. Дослідження властивостей критерію Пірсона зручно виконувати для статистичних даних, згрупованих на інтервалах з однаковою ймовірністю попадання на кожний з них. Для рівномірного закону розподілу генеральної сукупності ця вимога виконується автоматично, але і будь-який гіпотетичний закон можна розбити на інтервали з однаковою теоретичною ймовірністю, отже різної ширини. Зрозуміло, що використання змінної ширини інтервалів групування при візуальному аналізі розподілів мало прийнятне, але при перевірці гіпотез про закон розподілу воно є дуже корисним, бо дозволяє уникнути небажаних ситуацій, коли до якогось інтервалу групування попадає невелика кількість спостережень (m_i<5). Перевірка на великій кількості реалізацій (N>5·10^5) показала, що полігони розподілу значень $K_{розр}$ не залежать від алгоритму вибору довжини інтервалів розбиття. На рис.1 наведено полігони розподілу для вибірок об'ємом n=100 з інтервалами однакової і змінної довжини при кількості інтервалів групування k=5 (по вертикальній осі – відносні частота w_i відповідних значень $K_{розр}$, $\sum w_i = 1$).



1 – інтервали сталої довжини;
2 – інтервали з сталою ймовірністю.

Рисунок 1 – Порівняння полігонів з різними інтервалами групування

Таким чином, без втрати загальності можна аналізувати властивості

критерію Пірсона для сталих значень m_i, в тому числі для нерівномірних розподілів генеральної сукупності.

Нормальний закон визначається двома параметрами – математичним сподіванням a і середнім квадратичним відхиленням σ, показниковий закон розподілу визначається одним параметром λ. В табл.1 наведено границі інтервалів групування, які забезпечують однакові ймовірності попадання в інтервали при різній кількості інтервалів групування k.

Таблиця 1 – Границі інтервалів групування

k=5			k=6			k=7		
Нормал.	Показн.	Ймов.	Нормал.	Показн.	Ймов.	Нормал.	Показн.	Ймов.
a–3,0σ	0,00	0,200	a–3,0σ	0,00	0,167	a–3,0σ	0,00	0,143
a–0,842σ	0,223/λ	0,200	a–0,967σ	0,182/λ	0,167	a–1,068σ	0154/λ	0,143
a–0,253σ	0,511/λ	0,200	a–0,431σ	0,405/λ	0,167	a–0,566σ	0,336/λ	0,143
a+0,253σ	0,916/λ	0,200	a	0,693/λ	0,167	a–0,180σ	0,560/λ	0,143
a+0,842σ	1,609/λ	0,200	a+0,431σ	1,099/λ	0,167	a+0,180σ	0,848/λ	0,143
a+3,0σ	5,0/λ		a+0,967σ	1,792/λ	0,167	a+0,566σ	1,253/λ	0,143
			a+3,0σ	5,0/λ		a+1,068σ	1,946/λ	0,143
						a+3,0σ	5,0/λ	

Результати роботи. При знаходженні $K_{розр}$ теоретичні частоти m_i шукаються в припущенні слушності гіпотетичного закону, числові характеристики якого обчислюються за наявними вибірковими даними. Вважається [2], що більш коректним є використання числових характеристик, обчислених за згрупованими даними, хоч на практиці частіше застосовують числові характеристики, обчислені з використанням усіх спостережень (за всією вибіркою). Розглянемо властивості критерію згоди Пірсона саме для такого випадку.

Відомо, що критерій Пірсона має розподіл χ^2 („хі-квадрат”) з $r=k–l–1$ ступенями вільності, де l – кількість параметрів, які визначають гіпотетичний закон розподілу і обчислені за вибірковими даними [1]. Можна тільки зауважити, що величини n_i є дискретними, отже і значення $K_{розр}$ є дискретними, розподіл яких наближається до неперервного розподілу χ^2.

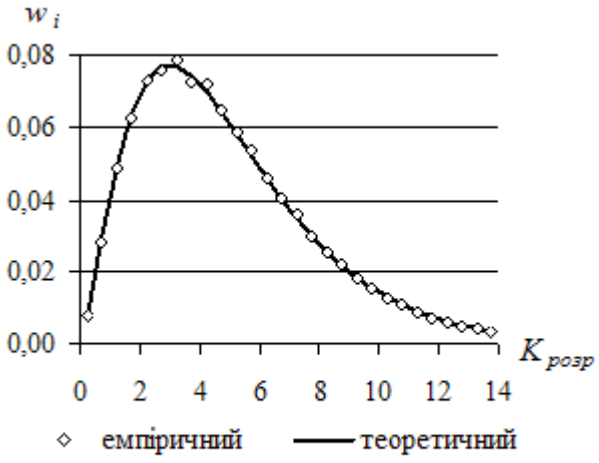


Рисунок 2 – Порівняння емпіричного і теоретичного розподілів

Якщо числові характеристики гіпотетичного закону розподілу задані апріорі, то l=0 і кількість ступенів вільності $r=k$-1. На рис.2 наведено емпіричний розподіл значень $K_{розр}$ для вибірок довжиною n=50 з кількістю інтервалів групування k=6 і заданими параметрами нормального закону (a=10, σ=3), і теоретичний розпо-

діл χ^2 з числом ступенів вільності r=5. Наведені дані свідчать про повну узгодженість графіків.

Розглянемо випадок, коли числові характеристики обчислюються за результатами вибірки, отже $l \neq 0$.

На рис.3 наведено порівняння емпіричного розподілу для вибірок довжиною n=100 з нормального закону розподілу, кількість інтервалів групування k=6, і теоретичні розподіли χ^2. Як видно на рисунках, емпіричний розподіл знаходиться між теоретичними кривими щільності розподілу χ^2 для числа ступенів вільності r=3 і r=4.

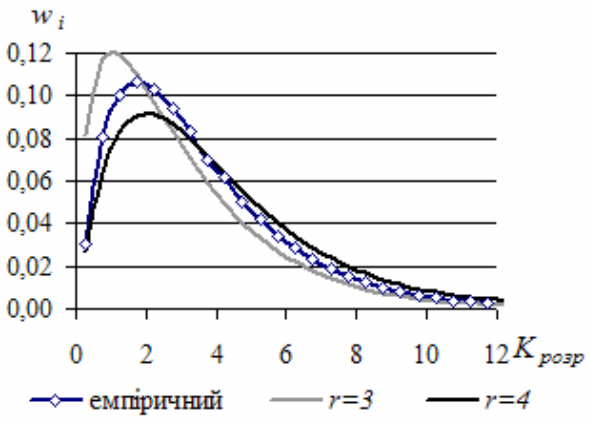


Рисунок 3 – Графіки емпіричного і теоретичних розподілів

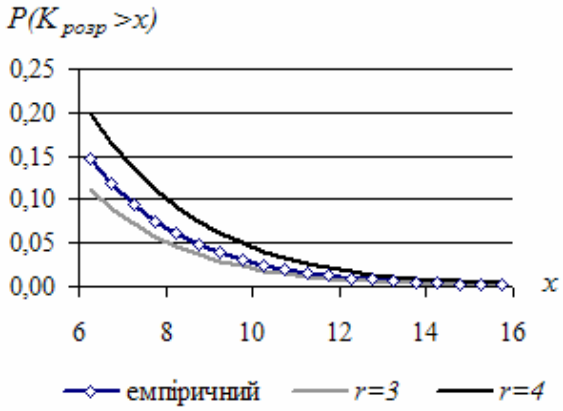


Рисунок 4 – Графіки критичних ймовірностей

Більш суттєва ця різниця проявляється на інтервалі знаходження критичних значень критерію Пірсона. Як відомо, рішення про прийняття чи відхилення гіпотези приймається при порівнянні розрахункових значень критерію з критичними, які знаходяться з умови $P\left(K_{\text{îïçä}} < K_{\text{êðèò}}\right) = \gamma = 1 - \alpha$, де γ – надійність висновків, α – рівень значущості (припустима ймовірність похибки першого роду звичайно приймається рівною 0,10; 0,05; 0,01).

На рис.4 показано графіки значень критичних ймовірностей для теоретичних і емпіричного розподілів, з якого можна зробити висновки, що емпіричний розподіл не зовсім відповідає теоретичним. Аналогічна ситуація виникає і при аналізі вибірок з кількістю інтервалів групування k=7.

В табл.2-4 наведено критичні значення $K_{крит}$ для емпіричних розподілів з нормально розподіленої генеральної сукупності і табличні значення $K_{крит}$ розподілу χ^2 при l=2. Для емпіричних розподілів критичні значення отримано з використанням експоненціальної апроксимації графіка критичних ймовірностей.

Таблиця 2 – Значення $K_{крит}$ для емпіричних і теоретичних розподілів, кількість інтервалів групування k=5

Рівень значущості α	Ступенів вільності r=2	Емпіричний розподіл	Ступенів вільності r=3
0,01	9,21	10,40	11,34
0,02	7,82	9,01	9,84
0,05	5,99	7,17	7,81
0,10	4,61	5,78	6,25

Таблиця 3 – Значення $K_{крит}$ для емпіричних і теоретичних розподілів, кількість інтервалів групування k=6

Рівень значущості α	Ступенів вільності r=3	Емпіричний розподіл	Ступенів вільності r=4
0,01	11,34	12,27	13,27
0,02	9,84	10,72	11,69
0,05	7,81	8,67	9,49
0,10	6,25	7,11	7,78

Таблиця 4 – Значення $K_{крит}$ для емпіричних і теоретичних розподілів, кількість інтервалів групування k=7

Рівень значущості α	Ступенів вільності r=4	Емпіричний розподіл	Ступенів вільності r=5
0,01	13,27	14,00	15,09
0,02	11,69	12,33	13,39
0,05	9,49	10,12	11,07
0,10	7,78	8,46	9,24

Порівняння полігонів розподілу значень $K_{розр}$ для вибірок різної довжини n показало деяку невідповідність їх інтуїтивним уявленням про степінь узгодженості реального і гіпотетичного розподілу.

Всі наведені дані стосуються випадку, коли числові характеристики вибірки обчислюються з використанням усіх її елементів, тобто за незгрупованими даними. В такому випадку критичні значення не залежать від довжини вибірки, а тільки від кількості інтервалів групування. На рис.5 наведено полігони значень критерію Пірсона для вибірок різної довжини n при кількості інтервалів групування k=6. Розподіли значень $K_{розр}$ співпадають в межах статистичних похибок, отже об'єм вибірки n суттєво не впливає на обґрунтованість висновків відносно припущення про закон розподілу генеральної сукупності. Дійсно, при перевірці гіпотези використовується тільки один параметр – кількість інтервалів групування k – і ніяким чином не враховується кількість використаних спостережень n.

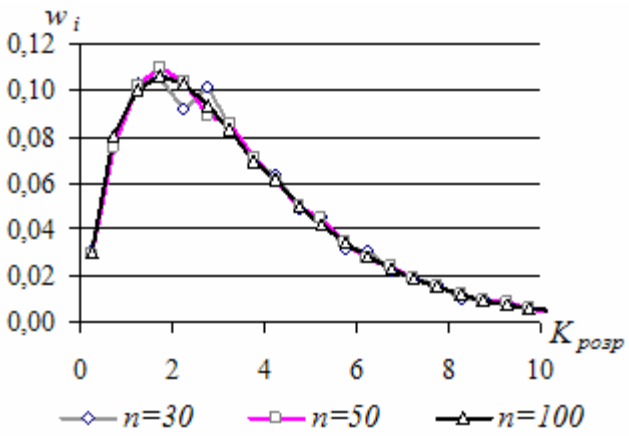


Рисунок 5 – Розподіли значень $K_{розр}$ для вибірок різної довжини n

На рис.6 наведено гістограми відносних частот для окремих вибірок з рівномірного закону розподілу різної довжини з кількістю інтервалів групування k=6. Для всіх вибірок значення розрахункового критерію однакові і дорівнюють $K_{розр}$=5,0. В той же час графіки свідчать, що відносні частоти для більших за об'ємом вибірок „в середньому" ближчі до теоретичного значення 0,167.

Такий неоднозначний зв'язок між величинами відхилу відносних частот від теоретичних і критерієм Пірсона витікає з виразу самого критерію.

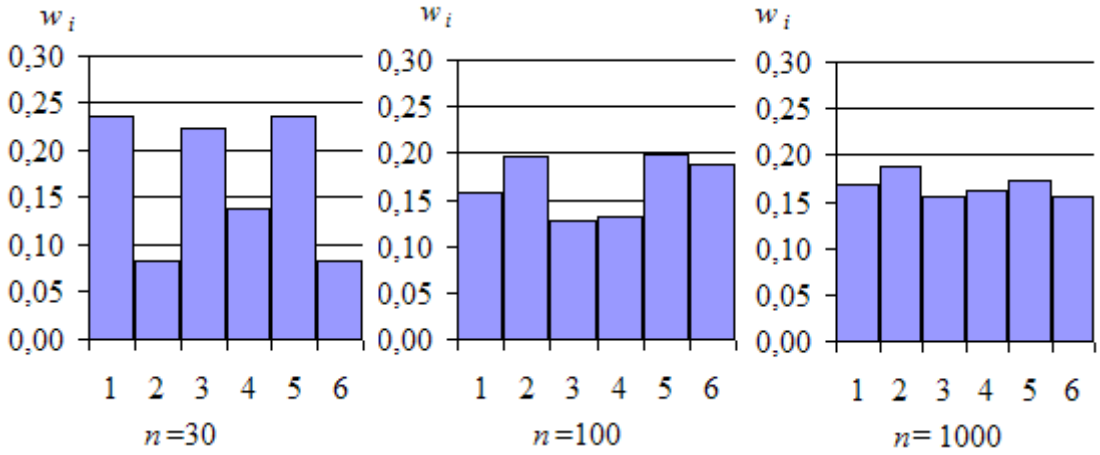


Рисунок 6 – Гістограми вибірок з однаковим значенням критерію Пірсона, $K_{розр}$ =5,0

Позначимо $m_i = n \cdot p_i$, $n_i = n \cdot w_i$, де p_i та w_i – відповідно теоретична ймовірність та емпірична відносна частота попадання до *i*-того інтервалу, *n* – об'єм вибірки. Тоді розрахункове значення критерію можна записати

$$K_{\partial i ç \partial} = \sum_{i=1}^{k} \frac{(m_i - n_i)^2}{m_i} = \sum_{i=1}^{k} \frac{(n \cdot p_i - n \cdot w_i)^2}{n \cdot p_i} = n \cdot \sum_{i=1}^{k} \frac{(p_i - w_i)^2}{p_i}.$$

Якщо використано розбиття на інтервали групування з однаковою імовірністю, то величина p_i є сталою і дорівнює $1/k$, отже

$$K_{\partial i ç \partial} = n \cdot p_i \cdot \sum_{i=1}^{k} \left(1 - \frac{w_i}{p_i}\right)^2 = n \cdot \frac{1}{k} \sum_{i=1}^{k} \left(1 - \frac{w_i}{p_i}\right)^2 = n \cdot \frac{1}{k} \sum_{i=1}^{k} \left(\frac{p_i - w_i}{p_i}\right)^2.$$

Вираз після першого множника *n* є середнім значенням квадрату відносних відхилень емпіричних відносних частот від теоретичних ймовірностей на відповідному інтервалі. Реальне зменшення цих відхилень при збільшенні об'єму вибірки, яке видно на рис.6, компенсується множенням на об'єм вибірки *n*, завдяки чому єдиним параметром при перевірці гіпотези залишається кількість інтервалів групування *k*.

Висновки. Таким чином, збільшення об'єму вибірки, що є гарантією наближення до значень числових характеристик генеральної сукупності, не є визначальним фактором при перевірці гіпотези про закон розподілу генеральної сукупності, бо кожна окрема вибірка, незалежно від її довжини, виступає як одна випадкова реалізація генеральної сукупності. Практична незалежність значень критерію Пірсона від об'єму вибірки дозволяє використовувати інший підхід для підвищення ефективності результатів перевірки слушності гіпотези про закон розподілу.

Велику за розмірами вибірку рекомендується розбити на кілька менших за розмірами і для кожної з них обчислити значення критерію Пірсона. За знайденими значеннями обчислюється середнє значення критерію $\overline{K}_{\partial i ç \partial}$. Кожен з доданків має χ^2 розподіл з *r* ступенями вільностей. Як відомо [3], їх сума також має розподіл χ^2, число ступенів вільностей якого дорівнює сумі ступенів вільностей доданків. Критичні значення розподілу середнього $\overline{K}_{\partial i ç \partial}$ можна шукати за таблицями критичних точок розподілу з відповідним числом ступенів вільностей, якщо результат поділити на кількість доданків.

На рис.7 наведено полігони емпіричних розподілів для вибірки *n*=100, а також для середніх з двох та трьох вибірок, кількість інтервалів групування *k*=6. У відповідності до теорії математичне сподівання середнього не міняється і дорівнює *r*, а диспер-

сія $\overline{K}_{\partial\hat{i}\varsigma\partial}$ зменшується і складає $2r/t$ (t – кількість вибірок для обчислення $\overline{K}_{\partial\hat{i}\varsigma\partial}$). На рис.8 наведено графіки емпіричних критичних ймовірностей для відповідних вибірок.

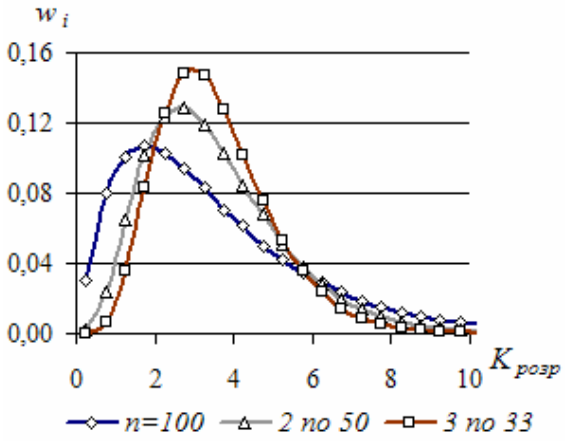


Рисунок 7 – Графіки емпіричних розподілів

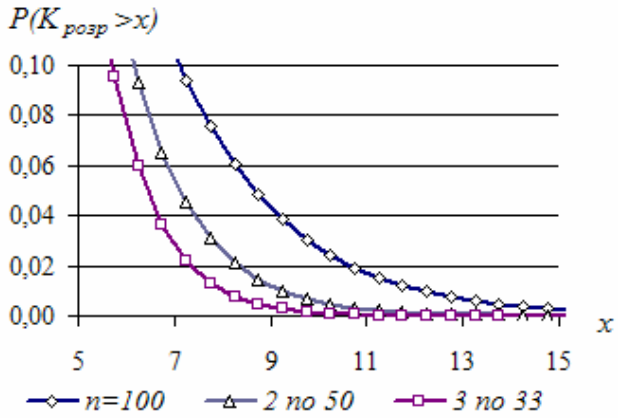


Рисунок 8 – Графіки критичних ймовірностей

Як і для окремих значень критерію $K_{розр}$, критичні значення для середніх $\overline{K}_{\partial\hat{i}\varsigma\partial}$ емпіричних дещо вищі від теоретичних. В табл.5 наведено критичні значення критерію Пірсона для емпіричних розподілів, отримані після апроксимації критичних ймовірностей, і теоретичні значення, отримані діленням табличних на кількість доданків.

Таблиця 5 – Значення $\overline{K}_{\hat{e}\partial\grave{e}\partial}$ для емпіричних і теоретичних розподілів, кількість інтервалів групування k=6

Рівень значущості α	$\overline{K}_{\hat{e}\partial\grave{e}\partial}$ – середнє з двох		$\overline{K}_{\hat{e}\partial\grave{e}\partial}$ – середнє з трьох	
	Емпіричний розподіл	Теоретичний r=6	Емпіричний розподіл	Теоретичний r=9
0,01	9,34	8,41	8,23	7,22
0,02	8,38	7,52	7,47	6,56
0,05	7,10	6,30	6,46	5,64
0,10	6,14	5,32	5,69	4,89

Використання усереднених розрахункових значень критерію $\overline{K}_{\hat{e}\partial\grave{e}\partial}$ і їх уточнених критичних значень дозволить більш обґрунтовано приймати рішення про відповідність законів розподілу вибіркової сукупності.

ЛІТЕРАТУРА

1 Жлуктенко В.І. Теорія ймовірностей і математична статистика: навч.-метод. посіб. у 2-х ч. – Ч. ІІ. Математична статистика / В.І.Жлуктенко, С.І.Наконечний, С.С.Савіна – К.: КНЕУ, 2001. – 336с.
2. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А.Айвазян, В.С.Мхитарян. – М.: ЮНИТИ, 1998. – 1022с.
3. Румшиский Л.З. Элементы теории вероятностей: учебник / Л.З.Румшиский. – М.: Наука, 1976. – 240с.

Надійшла до редколегії 03.02.2014.