ГОДИШНИК НА УНИВЕРСИТЕТА ПО АРХИТЕКТУРА, СТРОИТЕЛСТВО И ГЕОДЕЗИЯ СОФИЯ

Том Volume 57 2024 Брой Issue 2

ANNUAL OF THE UNIVERSITY OF ARCHITECTURE, CIVIL ENGINEERING AND GEODESY SOFIA

Получена: 01.09.2023 г.
Приета: 22.02.2024 г.

ЗНАЧЕНИЕ НА БРОЯ И ГРАНИЦИТЕ НА ИНТЕРВАЛИТЕ ПРИ АНАЛИЗ НА ГЕОДЕЗИЧЕСКИ ДАННИ ЗА ПРОВЕРКА НА ХИПОТЕЗАТА ЗА НОРМАЛНО РАЗПРЕДЕЛЕНИЕ НА ГРЕШКИТЕ ЧРЕЗ КРИТЕРИЯ НА ПИЪРСЪН

Т. Илиева[1], П. Павлов[2]

Ключови думи: *геодезически данни, проверка на хипотезата за нормалното разпределение, критерий на Пиърсън, χ^2 (хи-квадрат) тест, определяне на броя и границите на интервалите*

РЕЗЮМЕ

При обработката и анализа на геодезически измервания и функции от измерени величини се прави оценка на точността и проверка за съгласуваност на разпределението на грешките в съответствие с нормалното разпределение. Получените при анализ на данните разлики между фиксирани и измерени стойности се разглеждат като статистически редове от случайни грешки. Критерият на Пиърсън, известен още като χ^2 (хи-квадрат) тест, в комбинация с асиметрията и ексцеса на изследваните редове се използва за потвърждаване или отхвърляне на хипотезата за нормално разпределение.

Стойностите в редовете, които се изследват, трябва да бъдат минимум 30, за да може да бъде направена по-надеждна оценка. Те се групират в определен брой интервали при определени граници. Чрез критерия на Пиърсън се съпоставя реалният брой на стойностите във всеки интервал с теоретичния брой, определен на база вероятността за попадането им в него. Затова определянето на адекватни брой и граници на интервалите се оказва от съществено значение.

[1] Тамара Илиева, гл. ас. д-р инж., кат. „Геодезия и геоинформатика“, УАСГ, 411Р, бул. „Хр. Смирненски“ № 1, 1046 София, e-mail: ilieva_tamara@yahoo.com

[2] Павел Павлов, проф. д-р инж., кат. „Геодезия и геоинформатика“, УАСГ, 311Р, бул. „Хр. Смирненски“ № 1, 1046 София, e-mail: pgp@abv.bg

1. Въведение

Често в геодезическата практика се анализират и оценяват различни данни – координатни разлики, разлики между измерени и изчислени величини, несъвпадения в затворени фигури и други. Множеството от стойности, които се получават в процеса на оценка като разлики между функции от измерени величини и други известни техни значения се третират като статистически редове от случайни грешки и се прави проверка на хипотезата за нормалното им разпределение. В тези случаи се използват статистически критерии за изследване на съгласуваност между емпиричното и теоретичното разпределение.

На теория, случайните грешки при измерванията и функциите от тях се считат за случайни величини с нормално разпределение, затова е важно да се потвърди тази хипотеза при изследване. Пример за посочено в нормативната уредба изискване за проверка на хипотезата за нормално разпределение на данни при оценка на кадастрални планове е заложено в Приложение № 6 на Наредба № РД-02-20-5 от 15 декември 2016 г. за съдържанието, създаването и поддържането на кадастралната карта и кадастралните регистри [1].

В [2 и 3] е дадена последователността, чрез която се изследват редове от случайни грешки в геодезическата практика. Стъпките през които се преминава са следните:

- Проверка за наличие на груби грешки – изчислява се средната квадратна грешка m за реда, като се изключи максималната абсолютна стойност. Проверява дали в реда има стойности, които превишават $3m$ – ако има такива, то те се изключват.
- Проверка за наличие на систематични грешки – проверява се дали положителните и отрицателните стойности са приблизително еднакъв брой и дали средната аритметична стойност от всички грешки е близка до нула.
- Оценка на точността – определят се r (вероятна грешка) и s (средна аритметична грешка) и се проверява дали съотношението на r:m и s:m са близки до теоретичните такива.
- Проверка за нормално разпределение на грешките – тази проверка се изпълнява чрез групиране на стойностите в интервали (при доверителни вероятности съответстващи на 1σ, 2σ и 3σ – 0,683, 0,955 и 0,9973), с използване на критерия на Пиърсън, изчисляване на асиметрията и ексцеса на изследвания ред. Тук в някои случаи се включва и критерият на Колмогоров.

Във всички области на науката един от най-разпространените критерии за потвърждаване на съгласуваност на статистически данни с предполагаем закон на разпределение е критерият на Пиърсън (χ^2). Той спада към групата на непараметричните критерии, т.е., за да се приложи, изследваните данни не следва да имат определено разпределение, а се основава на честота или ранг.

Критерият χ^2 е предложен и публикуван от К. Пиърсън през 1900 г. за изследване във вероятностно отношение на наблюдавана система, като се определи годността ѝ чрез честотата на случайни извадки [4]. Стойностите в дадена извадка се групират в интервали и ако за наблюдаваната система стойността на χ^2, която представлява сума от повдигнатите на квадрат разлики между емпиричния и теоретичен брой стойности във всеки от интервалите, е много малка, то може честотното разпределение на данните да се счита като случайна вариация на предварително известен закон за разпределение. При това положение се счита, че произволната извадка характеризира генералната

съвкупност и се различава от нея само със случайни грешки в константите, изведени от наблюдаваната проба като емпиричен закон на разпределение.

За прилагане на критерия на Пиърсън (χ^2) по отношение на резултатите от геодезическите измервания или техни функции [2, 3] данните се третират като статистически ред и се групират в N на брой интервали. Ако в даден интервал броят на грешките е по малък от 5, то той се обединява със съседен такъв. Определят се нормираните граници на интервалите:

$$t_a = \frac{a - M(\varepsilon)}{m}, \quad t_b = \frac{b - M(\varepsilon)}{m}. \tag{1}$$

Определя се теоретичният брой на грешките във всеки интервал np_i. Изчислява се коефициентът χ^2 по формулата:

$$\chi^2 = \sum_{i=1}^{N} \frac{(k_i - np_i)^2}{np_i}. \tag{2}$$

Със стойността на χ^2 и степените на свобода r се определя вероятността $p(x^2, r)$:

$$p(x^2, r) = \frac{x^{\frac{r}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{r}{2}} \Gamma\left(\frac{r}{2}\right)}, \text{ където } \Gamma(n) = \int_0^{\infty} e^{-t} t^{n-1} dt. \tag{3}$$

Степените на свобода r се изчисляват по формулата:

$$r = N - l - 1, \tag{4}$$

където l е броят на неизвестните параметри на закона за разпределение. При нормално разпределение l = 2, но при изследване на случайни грешки l = 1, предвид това, че $M(\varepsilon) = 0$, т.е. остава само един параметър, който се оценява.

За определяне на оптималния брой на интервалите N при n на брой стойности се използва най-често формулата на Стърджис [5]:

$N \approx 1 + \log_2 n$, която може да се запише и във вида $N \approx 1 + 3{,}322 \log_{10} n$.

Тази формула, както и посочените по-долу такива, се използва също при групиране на данните за построяване на хистограми, но не се препоръчва за големи по обем извадки, тъй като при нея се апроксимира нормалното разпределение чрез биномни коефициенти (разпределение на Бернули) [6].

За определяне на броя на интервалите, в които стойностите ще се групират, може да се използват и други формули [7], например:

$N \approx 2\sqrt[3]{n}$ – формула на Райс (Rice, 1944),

$N \approx \sqrt{\frac{n}{5}}$ – формула на Кохран (Cohran, 1954),

$N \approx \sqrt[3]{2n}$ – формула на Терел и Скот (Terrel and Scott, 1985).

Има и други формули за определяне на приблизителния брой интервали за групиране на анализираните данни. Те се основават не само на броя стойности в

извадките, но и на размаха им и интерквартилния размах ($Q_3 - Q_1$). Може да се използва и разпределение по равно на броя стойности в интервалите. Тези методи обаче не са разгледани тук, предвид това, че са строго специфични за всеки изследван ред и не предполагат симетрия на данните във вероятностно отношение спрямо математическото очакване.

При проверката за потвърждаване или отхвърляне на хипотезата за нормално разпределение на данните броят стойности, които се изследват, е от значение. В геодезията обикновено се работи с извадки от по 30, 50, 100, 150, 200 или повече стойности. От значение е също и начинът, по който тези стойности се разпределят в интервалите, като обикновено се поставят допълнителни изисквания интервалите да са четен брой, разположени симетрично спрямо математическото очакване, а броят на грешките в тях да е не по-малък от 5.

При стойности на определената вероятност $p(\chi^2, r)$ от 0,1 до 0,3 се счита, че опитните данни не противоречат на хипотезата за нормално разпределение, а ако стойността е над 0,3, то хипотезата се потвърждава.

2. Методология на изследването и получени резултати

При настоящото изследване е определен приблизителен брой на интервалите за различни по брой стойности с използване на формулите по т. 1. Приблизително определяне на броя на интервалите със закръгление до цяло число с използване на формулите по т. 1 при различен обем на извадките е дадено в табл. 1. Направени са тестове за по-малките по обем извадки, тъй като в геодезическата практика по-рядко се обработват много големи обеми от данни, но на следващ етап изследването може да се разшири.

Таблица 1. Приблизително определяне на броя на интервалите при различен обем данни

Метод за приблизително определяне на броя на интервалите	**Брой стойности (обем на извадките)**											
	30	**50**	**100**	**150**	**200**	**300**	**400**	**500**	**1000**	**2000**	**5000**	**10000**
Метод на Стърджис	6	7	8	8	9	9	10	10	11	12	13	14
Метод на Райс	6	7	9	11	12	13	15	16	20	25	34	43
Метод на Кохран	2	3	4	5	6	8	9	10	14	20	32	45
Метод на Терел и Скот	4	5	6	7	7	8	9	10	13	16	22	27

От табл. 1 може да се види, че макар и стойностите да са близки, при различните обеми на данните в извадките може да се използва различен брой на интервалите. Напр. при извадка от 200 стойности може да се използва съответно 6, 8, 10 или 12 като брой на интервалите (избрани са само четни стойности, предвид това, че предварително се налага условието за четен брой). Стойността на χ^2 и броят на степените на свобода ще бъдат близки, но различни, като идеята тук е да се установи и това кога вероятността ще бъде най-висока и кога най-ниска.

Разгледани са различни по обем извадки и случаи на различен брой интервали при доверителна вероятност 0,9973 в граници от -3σ до +3σ. Резултатите са представени в табл. 2 до табл. 5. За изчисляване на вероятността p е използвана функция в Microsoft Excel, която връща вероятност при зададени стойности на χ^2 и r – CHIDIST(χ^2,r). Крайните интервали при -3σ и +3σ са приети с по-широки граници с цел да се избегне, доколкото това е възможно, обединяване на интервали със съседните им. При всяко обединяване на интервал със съседен такъв се намалява стойността на χ^2, а също се понижават и степените на свобода с единица.

Разгледан е примерът в [3], табл. 22.1, стр. 137, в който са дадени 104 стойности на несвръзки в триъгълници, т.е. определени са истински грешки. Направено е групиране на стойностите в различен брой интервали, а резултатите са представени в табл. 2.

Таблица 2. Критерий χ^2, приложен за 104 стойности

Граници на интервала		Бр. на инт.	Размер (ширина) на интервалите	χ^2	r – степени на свобода	ΣP_i	Вероятност
от	до						
-3σ	+3σ	6	1,5σ, 1σ, 0,5σ, 0,5σ, 1σ, 1,5σ	1,890	4	0,9973	75,595 %
-3σ	+3σ	8	1,5σ, 0,5σ, 0,5σ, 0,5σ, 0,5σ, 0,5σ, 0,5σ, 1,5σ	3,066	6	0,9973	80,052 %
-3σ	+3σ	10	1,5σ, 0,5σ, 0,5σ, 0,25σ, 0,25σ, 0,25σ, 0,25σ, 0,5σ, 0,5σ, 1,5σ	3,192	8	0,9973	92,145 %

Получената вероятност тук във всички случаи е по-голяма от 30 %, следователно хипотезата за нормално разпределение на опитните данни се потвърждава. При нарастване на броя на интервалите нарастват също стойностите на χ^2 и степените на свобода, а също и вероятността за потвърждаване на хипотезата за нормално разпределение.

Разгледан е същият пример, но с 50 стойности на несвръзки в триъгълници (първите 50 стойности). Резултатите са представени в табл. 3.

Таблица 3. Критерий χ^2, приложен за 50 стойности

Граници на интервала		Бр. на инт.	Размер (ширина) на интервалите	χ^2	r – степени на свобода	ΣP_i	Вероятност
от	до						
-3σ	+3σ	4	2σ, 1σ, 1σ, 2σ *	2,794	2	0,9973	42,450 %
-3σ	+3σ	6 => 5	1,5σ, 1σ, 0,5σ, 0,5σ, 1σ, 1,5σ **	2,269	4 => 3	0,9973	51,848 %
-3σ	+3σ	6 => 5	1,5σ, 0,75σ, 0,75σ, 0,75σ, 0,75σ, 1,5σ ***	1,574	4 => 3	0,9973	66,528 %

При групирането в 4 интервала се получават съответно 8, 14, 16, 12 стойности, а при групирането в 6 интервала при различни граници се получават съответно 1, 12, 9, 8, 15, 5 ** стойности и 1, 10, 11, 14, 9, 5 ***. При групирането в 6 интервала и в двата случая се налага обединяване на единия краен интервал със съседен, но предвид различните граници, и стойностите за χ^2* са различни.

Тъй като и тук получената вероятност във всички случаи е по-голяма от 30 %, то следва, че хипотезата за нормално разпределение на опитните данни се потвърждава. И тук се наблюдава увеличаване на вероятността при увеличаване на броя на интервалите, но след обединяването стойността на χ^2 и степените на свобода намаляват.

Разгледан е пример с 30 стойности на несвръзки в триъгълници (последните 30 стойности от табл. 3). Резултатите са представени в табл. 4.

Таблица 4. Критерий χ^2, приложен за 30 стойности

Граници на интервала		Бр. на инт.	Размер (ширина) на интервалите	χ^2	r – степени на свобода	ΣP_i	Вероятност
от	до						
-3σ	+3σ	4	2σ, 1σ, 1σ, 2σ *	1,615	2	0,9973	65,603 %
-3σ	+3σ	6 => 4	1,5σ, 0,75σ, 0,75σ, 0,75σ, 0,75σ, 1,5σ **	0,418	4 => 2	0,9973	81,154 %

При групирането в 4 интервала се получават съответно 5, 10, 8, 7 стойности, а при групирането в 6 интервала** се получават съответно 3, 4, 8, 7, 5, 3 стойности и се налага обединяване на двата крайни интервала със съседните им, като така пак се достига до 4 интервала, но с граници, различни от първия случай.*

И тук получената вероятност е над 30 %, следователно хипотезата за нормално разпределение на опитните данни се потвърждава и при двата варианта на групиране на стойностите.

Разгледан е и друг пример със 198 стойности. Резултатите са представени в табл. 5.

Таблица 5. Критерий χ^2, приложен за 198 стойности

Граници на интервала		Брой интервали	Размер (ширина) на интервалите	χ^2	r – степени на свобода	ΣP_i	Вероятност
от	до						
-3σ	+3σ	6	1,5σ, 1σ, 0,5σ, 0,5σ, 1σ, 1,5σ	6,781	4	0,9973	14,791 %
-3σ	+3σ	8	1,5σ, 0,5σ, 0,5σ, 0,5σ, 0,5σ, 0,5σ, 0,5σ, 1,5σ	8,308	6	0,9973	21,640 %
-3σ	+3σ	10	1,5σ, 0,5σ, 0,5σ, 0,25σ, 0,25σ, 0,25σ, 0,25σ, 0,5σ, 0,5σ, 1,5σ	9,881	8	0,9973	27,350 %
-3σ	+3σ	12 => 10	1σ, 0,5σ, 0,5σ, 0,5σ, 0,25σ, 0,25σ, 0,25σ, 0,25σ, 0,5σ, 0,5σ, 0,5σ, 1σ*	9,803	10 => 8	0,9973	27,916 %

При групирането в 12 интервала се получават съответно 4, 7, 10, 27, 20, 29, 19, 16, 32, 19, 11, 4, затова двата крайни интервала са обединени със съседните им.*

Тъй като тук получената вероятност във всички случаи е по-голяма от 10 %, но по-малка от 30 %, то следва, че опитните данни не противоречат на хипотезата за нормално разпределение.

3. Обобщение на получените резултати

Като обобщение може да се каже, че критерият χ^2 може да бъде управляван чрез начина на определяне на интервалите – техния брой и съответните им граници.

От направения анализ ясно се вижда, че при нарастване на броя на интервалите се повишават степените на свобода, съответно и вероятността за потвърждаване на хипотезата за нормално разпределение. Трябва да се има предвид обаче, че не е желателно изкуствено скъсяване на ширината на интервалите и увеличаване на броя им,

тъй като намалява и броят на грешките в тях. Не е желателно и изкуствено разширяване на границите на интервалите, напр. чрез обединяване, тъй като и това повлиява на стойността на χ^2.

На базата на представените резултатите от проверка на хипотезата за нормално разпределение на различни по обем извадки от опитни данни за до 200 стойности може да се обобщи следното:

- При извадки с обем от порядъка на 30 – 50 стойности по-добре е данните да се групират в 4 интервала, с по-голяма ширина на крайните такива. Тук формулата на Терел и Скот дава именно такъв брой на интервалите, докато с тази на Стърджис се получават 6 интервала, при които почти винаги се налага обединяване на крайните интервали.

- При извадки с обем от порядъка на 100 до 200 стойности е по-добре да се групират данните съответно в 6 или 8 интервала, с по-голяма ширина на крайните такива, предвид това да има достатъчен брой стойности във всеки интервал. Тук приблизителният брой стойности може да се определи както по формулата на Терел и Скот, така и с тази на Стърджис.

ЛИТЕРАТУРА

1. Prilozhenie № 6 kam chl. 82, al. 6 (Analiz na statisticheski red. Otsenka na tochnostta. Sadarzhanie na doklada za otsenka na tochnostta na kadastralen plan) na Naredba № RD-02-20-5 ot 15 dekemvri 2016 g. za sadarzhanieto, sazdavaneto i poddarzhaneto na kadastralnata karta i kadastralnite registri (posl. izm. i dop. DV. br.72 ot 31 avgust 2021 g.), https://www.cadastre.bg/sites/default/files/documents/regulation/13994/naredba-rd-02-20-5-ot-2016-g.pdf.

2. *Atanasov, S*. Teoria na matematicheskata obrabotka na geodezicheskite izmervania. S., Tehnika, 1988, https://uacg.bg/UserFiles/File/FGE/BOOKS/TMOGI(1).pdf.

3. *Kostadinov, K., Valchinov, V*. Matematicheska obrabotka na geodezicheski izmervania. S., UASG, 2012.

4. *Karl Pearson, F.R.S*. (1900): X. 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 Series 5, 50:302, 157 – 175, http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/Proportion/Pearson1900.pdf.

5. *Sturges, H*. The choice of a class-interval. J. American Statist. Assoc., 21, 65 – 66, 1926.

6. *Hyndman, Rob*. The problem with Sturges' rule for constructing histograms. Department of Econometrics and Business Statistics, 1995, https://www.researchgate.net/publication/222105804_The_problem_with_Sturges'_rule_for_constructing_histograms#fullTextFileContent.

7. *Nurhan Doğan, Doğan, I*. Determination of the number of bins/classes used in histograms and frequency tables: a short bibliography. TurkStat, Journal of Statistical Research, December 2010, https://dergipark.org.tr/tr/download/article-file/2579010.

SIGNIFICANCE OF INTERVAL NUMBER AND LIMITS IN THE ANALYSIS OF GEODETIC DATA FOR TESTING THE HYPOTHESIS OF NORMAL ERROR DISTRIBUTION BY PEARSON'S CRITERION

T. Ilieva[1], P. Pavlov[2]

Keywords: *geodetic data, normal distribution hypothesis testing, Pearson's criterion, χ^2 (chi-squared) test, interval number and limits*

ABSTRACT

In the processing and analysis of geodetic measurements and functions of measured values, an assessment of accuracy and a check for consistency of error distribution in accordance with the normal distribution are performed. Differences between fixed and measured values obtained during data analysis are considered as statistical series of random errors. Pearson's criterion, also known as χ^2 (chi-squared) test, in combination with skewness and kurtosis of the studied series is used to confirm or reject the hypothesis of normal distribution.

The values in the examined series should be minimum 30 for a more reliable estimate to be performed. They are grouped into intervals with certain number and limits. Pearson's test compares the real number of values in each interval with the theoretical number determined on the basis of the probability for it. Therefore, the determination of adequate interval number and limits turns out to be essential.

[1] Tamara Ilieva, Chief Assist. Prof. Dr. Eng., Dept. "Geodesy and Geoinformatics", UACEG, 1 H. Smirnenski Blvd., Sofia 1046, e-mail: ilieva_tamara@yahoo.com

[2] Pavel Pavlov, Prof. Dr. Eng., Dept. "Geodesy and Geoinformatics", UACEG, 1 H. Smirnenski Blvd., Sofia 1046, e-mail: pgp@abv.bg