УДК 004.42: 519.22

Білоконь Алла, ст. магістратури факультету кібернетики; науковий керівник – д ф.-м.н., професор Джунь Й. В. (Міжнародний економіко-гуманітарний університет імені академіка Степана Дем'янчука, м. Рівне)

УМОВИ ЗАСТОСУВАННЯ ОДНОФАКТОРНОГО ДИСПЕРСІЙНОГО АНАЛІЗУ ПРИ ВИКОРИСТАННІ ПРОГРАМНИХ ПРОДУКТІВ «ANOVA»

Анотація. *У статті досліджено попередню перевірку фундаментальних математичних принципів дисперсійного аналізу при обробці результатів наукових чи виробничих експериментів перед застосуванням програмного продукту «ANOVA», для перевірки гіпотези рівності дисперсій запропоновано M-статистику Бартлетта, а для перевірки нормальності вибірок – d-статистику. Для забезпечення коректного застосування «ANOVA», запропоновано розробити додатковий програмний модуль, який рекомендовано застосовувати перед запуском програми.*
Ключові слова: *дисперсійний аналіз, критерій Бартлетта, d-статистика.*

Аннотация*. В статье исследована предварительная проверка фундаментальных математических принципов дисперсионного анализа при обработке результатов научных или производственных экспериментов перед применением программного продукта «ANOVA», для проверки гипотезы равенства дисперсий предложена M-статистика Бартлетта, а для проверки нормальности выборок – d-статистика. Для обеспечения корректного применения «ANOVA», предложено разработать дополнительный программный модуль, который рекомендуется применять перед запуском программы.*
Ключевые слова: *дисперсионный анализ, критерий Бартлетта, d-статистика.*

Annotation. *The article shows the importance of the preliminary examination of fundamental mathematical principles of analysis of variance in the processing of the results of scientific and industrial experiments before application software product «ANOVA». These principles are: equality of variances and the normal output samples. To test the hypothesis of equality of variances it was suggested Bartlett M-statistics, and to verify the normality of sampling – d-statistics. To ensure the correct application of «ANOVA», it is proposed to develop additional software module that is recommended before starting the program.*
Keywords*: analysis of variance, Bartlett criterion, d-statistics.*

Інтелектуальний аналіз даних (ІАД) широко застосовується для методів теорії ймовірностей і математичної статистики, а також процедур

математичного моделювання для отримання необхідних висновків наукового, виробничого, маркетингового чи соціального характеру [1-5]. Одною з важливих аналітичних складових ІАД є дисперсійний аналіз (ДА), який обслуговується програмним продуктом «ANOVA». Проте, застосування цього продукту до реальних даних не завжди є коректним, тому що математичні положення, покладені в основу ДА, часто не відповідають реальній практиці спостережень. Наприклад, однією з основних умов застосування ДА є нормальність розподілу його вибірок.

Проте, масова перевірка цієї гіпотези, здійснена Лабораторією прикладної математики Тартуського державного університету показала, що з 2500 вибірок реальних статистичних даних 92 % є не гаусовими [5]. Більш детальні відомості з цього питання наведені в роботах [6–8; 11] у яких вже розглянуті випадки, коли навіть 100 % вибірок виробничого характеру не є гаусовими.

На думку сучасних науковців, коректне застосування ДА до таких даних є неможливим. Проте, в існуючому програмному продукті «ANOVA» не передбачена перевірка самих умов застосування ДА, оскільки вважається, що програмісти в достатній мірі кваліфіковані і розуміють суть цих умов. Насправді ж ці умови користувачі програмних продуктів не перевіряють, вважаючи, що автори цих програмних засобів в усьому розібрались. Це і є головною причиною непрофесійного використання програмних продуктів «ANOVA» [1; 2; 3].

Метою нашої статті є дослідження та розроблення рекомендацій перевірки умов застосування «ANOVA» для реальних емпіричних даних. Ці рекомендації можуть бути основою для розроблення спеціального програмного додатку до «ANOVA», який би забезпечував високий професіоналізм роботи кожного програміста при математичній обробці реальних статистичних даних наукового чи виробничого характеру.

Для перевірки умов застосування «ANOVA» рекомендується використовувати два критерії:

– Бартлетта – для перевірки гіпотези про рівність дисперсій у вибірках при умові, коли $k > 2$;

– d-статистику для перевірки гіпотези нормальності вибірок,

тому що це найбільше ефективний, простий і універсальний засіб перевірки для вибірок обсягом від 10 до 1000 спостережень.

Перевірка умови рівності дисперсій вибірок в ДА зводиться до обчислення *M*-статистики:

$$M = N ln\left(\frac{1}{N}\sum_{i=1}^{k} \vartheta_i S_i^2\right) - \sum_{i=1}^{k} \vartheta_i ln S_i^2 \; ; \qquad (1)$$

де S_1^2, S_2^2, …, S_k^2 взаємонезалежні незміщені статистичні оцінки дисперсій 6_1^2, 6_2^2, …, 6_k^2; $k -$ число вибірок, $\vartheta_i = n_i - 1$, n_i – обсяг i-тої вибірки;

$$N = \sum_{i=1}^{k} \vartheta_i.$$

Якщо гіпотеза H_0 ($6_1^2 = 6_2^2 = \ldots = 6_k^2 = 6^2$) вірна і усі $\vartheta_i > 3$, то відношення:

$$M \left[1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{\vartheta_i} - \frac{1}{N}\right)\right]^{-1};$$

розподілено приблизно як χ^2 з $k-1$ ступенями свободи [9, с.46].

Зауважимо також, що M-критерій Бартлетта є досить чутливим до відхилень істинних розподілів величин $\vartheta_i S_i^2 / 6^2$ від χ^2 розподілу Пірсона.

Зокрема, якщо всі оцінки S_i^2 побудовані по вибіркам із сукупностей, розподіли яких відмінні від нормальних, то М-критерій може з великою імовірністю відхилити гіпотезу H_0, коли вона є вірною.

Після отримання M-статистики по формулі (1), критерій Бартлетта зводиться до перевірки альтернативи:

$$M < M_{\text{кр},\alpha,k,c_1}, \quad (2)$$

$$M \geq M_{\text{кр},\alpha,k,c_1}, \quad (3)$$

де α – рівень значущості критерію (ризику); k – кількість вибірок, рівність дисперсій яких перевіряється:

$$c_1 = \sum_{i=1}^{k} \vartheta_i - N^{-1}; \quad (4)$$

– критичне значення M-статистики, яке вибирається з таблиць [9, с. 239] в в залежності від α, k і c_1.

Якщо виконується нерівність (2), то це означає що дисперсії у всіх досліджуваних вибірках рівні. При виконанні відношення (3) гіпотеза H_0 відхиляється, а це означає, що дисперсії S_i^2 істотно відрізняються між собою.

Як було зазначено раніше, для застосування *M*-критерію і ДА, надзвичайно важливою є перевірка нормальності розподілу вибірок. Ця перевірка здійснюється за допомогою d-статистики [9, с. 55]:

$$\boldsymbol{d = \frac{1}{nS} \sum_{l=1}^{n} |x_l - \bar{x}|,} \quad \textbf{(5)}$$

де x_l результат вибіркових спостережень (l=1,2,…,n); вибіркове середнє $\bar{x} = \sum_{l=1}^{n} x_l/n$; вибіркова дисперсія $S^2=\sum_{l=1}^{n}(x_l - \bar{x})^2/n$.

В ідеальному випадку для істинно розподілених випадкових величин ξ, що мають математичне сподівання a і дисперсію б:

$$\boldsymbol{d = \frac{M|\xi - a|}{б} = \sqrt{\frac{2}{\pi}} = 0.79788,}$$

де M – оператор математичного сподівання.

Перевірка гіпотези нормальності за допомогою d-статистики зводиться до визначення справедливості однієї з наступних нерівностей:

$$\boldsymbol{d_{5\%} < d < d_{95\%};} \quad \textbf{(6)}$$

$$\boldsymbol{d_{95\%} \geq d;} \quad \textbf{(7)}$$

$$\boldsymbol{d \leq d_{5\%};} \quad \textbf{(8)}$$

де – відповідно 5 % і 95 % квантилі розподілу d-статистики, які вибираються з таблиць [9, с. 258]

Значення d обчислюються для кожної вибірки. Якщо обчислене d попадає в межі (6), то це свідчить про нормальність вибірки. В тому випадку, коли виконуються відношення (7; 8), – гіпотеза нормальності відхиляється з 10 % ризиком.

Застосування програмного продукту «ANOVA» є коректним лише в тому випадку, коли критерій Бартлетта підтвердить рівність дисперсій, а значення d-статистики підтвердить виконання гіпотези нормальності для кожної вибірки. Тобто, коли підтверджуються нерівність (3) і нерівність (6) для кожної вибірки. Коли буде в наявності хоча б один випадок не підтвердження, то це буде означати, що наявні дані є не коректними для того, щоб успішно застосувати ДА. Ігнорування такими перевірками може приводити до тяжких, а іноді і до катастрофічних наслідків. Прикладом цього є катастрофи крупних американських інвестиційних корпорацій LTKM у 1998 р. і «Amarant» у 2006 р., які в своїй роботі використовували

неадекватні методики моделювання і не перевірений реальний закон похибок математичної моделі [8; 10].

1. Ліщина Н. М. Інтелектуальний аналіз даних : конспект лекцій для студентів напряму підготовки «Комп'ютерні науки» / уклад. Н. М. Ліщина. – Луцьк : Луцький НТУ, 2016. – 112 с. **2.** Андрущак І. Є. Системний аналіз : конспект лекцій для студентів спеціальності 121 «Інженерія програмного забезпечення». / І. Є. Андрущак. – Луцьк : Луцький НТУ, 2014. – 84 с. **3.** Барсегян А. А. Технологии анализа данных : Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – 2-е изд., перераб. и доп. – СПб. : БХВ-Петербург, 2007. – 384 с. **4.** Tang Z., MacLennan J. Data Mining with SQL Server 2005. – Wiley Publishing, Inc., Indianapolis, Indiana. 2005. – 483с. **5.** Тоодинг Л. М. Проверка нормальности реальных статистических выборок. Материалы IV Всесоюзной научно-технической конференции «Применение многомерного статистического анализа в экономике и оценке качества продукции». Тезисы докладов / Тоодинг Л. М. – Тарту : ТГУ, 1989. – С. 262 – 263. **6.** Орлов А. И. Часто ли распределение результатов наблюдений является нормальным? / А. И. Орлов. – Заводская лаборатория. – 1971. – № 7. – С. 64 – 66. **7.** Новицкий П. В. Оценка погрешностей результатов измерений. – 2-е изд., перераб. и доп. / П. В. Новицкий. – Л. : Энергоатомиздат. Ленингр. отделение, 1991. – 304 с. **8.** Джунь И. В. Неклассическая теория погрешностей измерений / И. В. Джунь. – Ривне : Изд. дом ЭСТЕРО, 2015. – 171 с. **9.** Большев Л. Н. Таблицы математической статистики / Л. Н. Большев– М. : Наука. Редакция физико-математической литературы, 1983. – 416 с. **10.** Талеб Н. Черный лебедь. Под знаком непредсказуемости / Талеб Н. – М. : Колибри, 2012. – 528 с. **11.** Jeffreys H. Theory of Probability / Jeffreys H. – Oxford : Clarend Press. – 1998. – 470 p.