УДК 519.237.5: 621.9

С.М. ЛАПАЧ[*]

РИЗИКИ ЗАСТОСУВАННЯ КОЕФІЦІЄНТА КОРЕЛЯЦІЇ ПРИ КОНКРЕТНІЙ СПЕЦИФІКАЦІЇ РЕГРЕСІЙНОЇ МОДЕЛІ

[*]Національний технічний університет України «КПІ», м. Київ, Україна

Анотація. *Розглядається проблема надійності вибіркового коефіцієнта кореляції при використанні його для визначення конкретної специфікації регресійної моделі. При конкретній специфікації визначається список членів моделі, які забезпечують найкращий бажаний набір вибраних характеристик моделі. Найчастіше для цього використовується одна навчальна вибірка, але можливий варіант із навчальною і екзаменаційною вибірками. В попередніх роботах показано, що найбільш надійним для визначення специфікації є використання коефіцієнта кореляції Пірсона. Разом з тим питання про межі його надійного застосування не розглядались. Необхідність такого дослідження викликана використанням у множинному регресійному аналізі малих коефіцієнтів кореляції, які формально можуть бути статистично не значимими. При цьому виникає два питання. Перше: яка абсолютна величина коефіцієнта кореляції може вважатись надійною для прийняття рішення про включення регресора в модель. Друге: який розмір має бути для екзаменаційної вибірки для того, щоб коефіцієнти кореляції в навчальній і екзаменаційній вибірках можна було віднести до однієї генеральної сукупності. Показано, що виділення за коефіцієнтом кореляції при значенні менше 0,2 сумнівне і необґрунтоване. В такій ситуації, по-перше, необхідні додаткові перевірки обґрунтованості включення таких регресорів у модель, а, по-друге, урахування цього факту при аналізі і використанні моделі. Перше, як правило, виконується, так як процедури вибору окремої структури зазвичай багатоступінчасті. Друге фактично забезпечується тим, що при експоненціальній формі розподілу сили впливу регресорів на «хвіст» моделі припадає незначна частина частки впливу, яка пояснюється моделлю. В тому випадку, коли матриця експерименту є результатом пасивного експерименту або випадковою, то при використанні екзаменаційної вибірки бажано використовувати для прийняття рішення медіану коефіцієнта кореляції, яка розрахована методом «складного ножа» або шляхом формування випадкових підвибірок з вихідної. Розмір екзаменаційної вибірки при застосуванні для конкретної специфікації двох вибірок може бути 0,25 від навчальної тільки в тому випадку, коли вона не менше 30 експериментів. В інших випадках необхідно приймати не менше, ніж 0,5.*

Ключові слова: *регресійний аналіз, вибірковий коефіцієнт кореляції, метод «складного ножа», конкретна специфікація регресійної моделі.*

Аннотация. *Рассматривается проблема надежности выборочного коэффициента корреляции для определения конкретной спецификации регрессионной модели. При конкретной спецификации определяется список членов модели, которые обеспечивают желательное множество выбранных характеристик модели. Чаще всего для этого используется только обучающая выборка, но возможны варианты с обучающей и экзаменационной. В предыдущих работах показано, что наиболее надежным для определения спецификации есть использование коэффициента корреляции Пирсона. Вместе с тем вопрос о границах его надежного использования не рассматривался. Необходимость такого исследования вызвана использованием во множественном регрессионном анализе малых по абсолютному значению коэффициентов корреляции, которые формально могут быть незначимыми. При этом возникают два вопроса. Во-первых, какая абсолютная величина коэффициента корреляции может считаться надежной для принятия решения о включении регрессора в модель. Во-вторых, каким должен быть размер экзаменационной выборки, чтобы коэффициенты корреляции обучающей и экзаменационной можно было отнести к одной генеральной совокупности. Показано, что выделение по коэффициенту меньше 0,2 сомнительное и необоснованное. В этой ситуации, во-первых, нужны дополнительные проверки обоснованности включения регрессоров в модель, а, во-вторых, учет этого факта при анализе и использовании модели. Первое, как правило, выполняется, поскольку процедуры формирования структуры, как правило, многоступенчатые. Второе фактически обеспечивается тем, что при экспоненциальной форме распределения силы влияния регрессоров на «хвост» модели падает незначительная часть доли влияния, которая*

объясняется моделью. В том случае, когда матрица эксперимента является случайной, результатом пассивного эксперимента, то при использовании экзаменационной выборки, желательно для принятия решения использовать медиану коэффициента корреляции, рассчитанную методом «складного ножа» или путем формирования множества случайных выборок из исходной. Размер экзаменационной выборки при использовании её для спецификации может быть равен 0,25 от обучающей только в случае, когда она не меньше 30 экспериментов. В противном случае её размер необходимо устанавливать не менее 0,5.
Ключевые слова: *регрессионный анализ, выборочный коэффициент корреляции, метод «складного ножа», частная структура регрессионной модели.*

Abstract. *The reliability problem of the selective correlation coefficient is considered to determine a certain specification of the regression model. With a certain specification, a list of model members is determined that provide the desired set of selected characteristics of the model. Most often, only a training sample is used for this, but variants with training and examinations are possible. In previous works it was shown that the use of the Pearson correlation coefficient is the most reliable for determining the specification. At the same time, the question of the limits of its reliable usage was not considered. The need for such an investigation is caused by the use of multiple correlations in multiple regression analysis, which can formally be insignificant. This raises two questions. First, which absolute value of the correlation coefficient can be considered reliable for making a decision to include a regressor in the model. Secondly, what should be the size of the examination sample, so that the correlation coefficients of the teaching and examination can be attributed to one general population. It is shown that the selection at a coefficient of less than 0,2 is doubtful and unreasonable. In this situation, firstly, additional checks are needed for the validity of the inclusion of regressors in the model, and, secondly, consideration of this fact in the analysis and use of the model. The first, as a rule, is carried out, as the procedures for the formation of the structure, as a rule, are multistage. The second is actually ensured by the fact that with an exponential form of distribution of the force of influence of the regressors, an insignificant part of the influence of the influence falls on the "tail" of the model, which is explained by the model. In the case where the experimental matrix is random, the result of a passive experiment or using an examination sample is desirable for making a decision to use the median of the correlation coefficient calculated by the «folding knife» method or by forming a set of random samples from the original one. The size of the examination sample when using it for the specification can be equal to 0,25 from the training sample only if it is not less than 30 experiments. Otherwise, its size should be set at least 0,5.*
Keywords: *regressive analysis, selective coefficient of correlation, method of «folding knife», private structure of regressive model.*

1. Вступ

Традиційно в регресійному аналізі (РА) використовуються дві вибірки: навчальна і контрольна. За навчальною вибіркою виконується ідентифікація моделі, а за контрольною – незміщена перевірка моделі на адекватність. Ці вибірки не повинні перетинатись. Фактично, на навчальній вибірці також виконується і остаточна специфікація моделі – визначення конкретної структури моделі (переліку регресорів, які входять у модель) [1, 2]. За контрольною визначається незміщена оцінка залишкової дисперсії для перевірки адекватності моделі за критерієм Фішера і оцінюються її прогностичні властивості.

О.Г. Івахненком у [3] висунута ідея трьох вибірок. Третьою є екзаменаційна, яка сумісно з навчальною використовується для специфікації моделі, що повинно підвищити надійність вибору елементів структури моделі.

Як показник для формування структури моделі будемо розглядати коефіцієнт парної кореляції Пірсона, використання якого обґрунтовано в [2, 4].

Малі значення коефіцієнта кореляції при звичайних статистичних дослідженнях вважаються статистично не значимими і не розглядається, але це не так в регресійному аналізі. В ньому часто в модель включаються фактори, коефіцієнт кореляції якого з відгуком досягає всього-на-всього 0,05 [1, 5]. Це пов'язано з цілим рядом особливостей регре-

сійного аналізу щодо формування структури рівняння регресії, розгляд якого виходить за межі даної роботи.

2. Основна частина

Виникає питання, який розмір повинні мати екзаменаційна і контрольна вибірки порівняно з навчальною. У прикладників склалась емпірична традиція, за якою екзаменаційна має розмір, який не перевищує 0,25, а контрольна – 0,1 від розміру навчальної. Якщо врахувати типовий розмір навчальної в 27–32 експерименти, то реально екзаменаційна має розмір 7–8 експериментів, а контрольна 3–5. А чи достатній такий розмір для надійного (обґрунтованого) прийняття рішення про структуру моделі чи її адекватність.

Крім того, незрозуміло, наскільки обґрунтованим є вибір конкретної специфікації моделі при малих значеннях коефіцієнта кореляції.

Мета статті: дослідити надійність прийняття рішень відносно структури моделі при різних рівнях залежності між регресорами і відгуком і визначити необхідний розмір екзаменаційної вибірки.

Вважається, що передумова гомоскедастичності виконується, а викиди відсутні.

Ті експериментатори, які приймають вказані вище розміри для екзаменаційної вибірки, не враховують той факт, що чим менше розмір вибірки, тим більша ймовірність відхилення емпіричних характеристик випадкової величини, визначених за цією вибіркою, від теоретичних [6]. Особливо це стосується таких малих вибірок, які розглядаються в даній роботі і використовуються в регресійному аналізі. Отриманий на випадковій вибірці з генеральної сукупності коефіцієнт кореляції є випадковою величиною, яка може відрізнятись від значення його в генеральній сукупності. А при прийнятті рішень у регресійному аналізі використовується просто вибіркове значення без будь-яких інших характеристик (довірчий інтервал, значимість тощо). У зв'язку з цим виникає питання про надійність прийняття рішень з приводу конкретної специфікації при відсутності екзаменаційної вибірки.

Розглянемо, як може змінюватись фактичне розраховане значення коефіцієнта кореляції при зміні розміру вибірки і тісноти зв'язку.

Довірчий інтервал для коефіцієнта кореляції визначається за формулою (1)

$$\Delta_r = \frac{t_{N-2,p}(1-r^2)}{\sqrt{N}}. \qquad (1)$$

На рис. 1 показані значення довірчих інтервалів для коефіцієнта кореляції при різних розмірах вибірки в порівнянні з можливим значенням цього коефіцієнта.

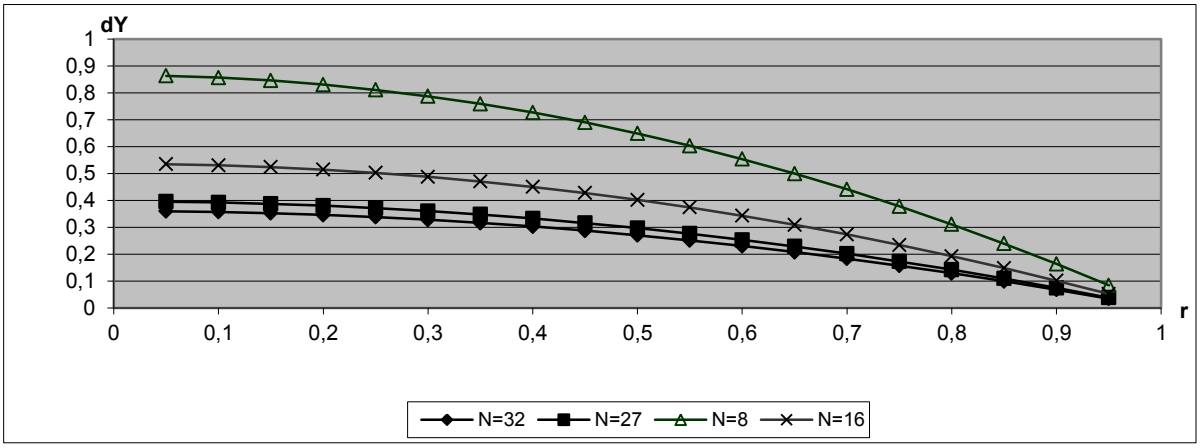

Рисунок 1 – Довірчі інтервали для коефіцієнта кореляції при різному розмірі вибірки

Можна зробити висновок, що при зазначених вище малих коефіцієнтах кореляції зв'язок таких регресорів з відгуком є статистично не значимим і їх не потрібно включати в модель. Але в [5] зазначається, що окремі регресори можуть бути не значимими, але їх сукупність значима і повинна входити в регресійну модель. В [7] показано, що «істинні» фактори мають більший коефіцієнт кореляції з відгуком, ніж мультиколінеарні з ними «хибні». Це викликає необхідність у порівнянні величин закорельованості регресорів з відгуком між собою для виконання вибору, незалежно від їх статистичної значимості. Практика побудови регресійних моделей дає приклад і інших ситуацій, коли в модель включають фактично статистично не значимі фактори. В [8] незначимі фактори включають у модель, тому що у теорії і практиці даної галузі знань вони вважаються впливаючими на відгук.

Для дослідження було проведено обчислювальний експеримент. Використовується спеціально створений приклад, в якому як фактори служать рівномірно розподілені в багатофакторному просторі псевдовипадкові $ЛП_\tau$ числа [9]. Відгуком є лінійна залежність різної тісноти зв'язку, «зашумлена» випадковими числами з нормальним законом розподілу. Вибірка має розмір 64. З неї випадковим чином вибираються по 50 підвибірок розміром 32, 18 і 8 відповідно, для яких проводиться аналіз варіації емпіричного коефіцієнта кореляції. Табл. 1–3 містять результати цього аналізу при різних рівнях залежності між змінними. Велика зміна значення кореляції свідчить про те, що цей експеримент належить до іншої генеральної сукупності. Таке може бути як з одним експериментом, так і з підмножиною. Цей випадок у роботі не розглядається.

Добре видно, що для сильної залежності (табл. 1) при зменшенні розміру вибірки розмах можливих значень збільшується, але знак розрахованих вибіркових коефіцієнтів кореляції не змінюється.

Таблиця 1 – Характеристики варіації коефіцієнта кореляції при сильній залежності

Оцінюваний параметр	Розмір вибірки		
	32	16	8
Мінімім	0,766612	0,680296	0,282838
Максимум	0,908861	0,945325	0,974052
Середнє	0,837239	0,848741	0,821998
Медіана	0,838949	0,861168	0,847246
Коефіцієнт кореляції на сукупності	0,843617		

При помірній залежності (табл. 2) при зменшенні розміру вибірки можлива зміна знаку вибіркових коефіцієнтів.

Таблиця 2 – Характеристики варіації коефіцієнта кореляції при помірній залежності

Оцінюваний параметр	Розмір вибірки		
	32	16	8
Мінімім	0,139179	-0,15465	-0,3269
Максимум	0,538873	0,731417	0,891748
Середнє	0,34313	0,309219	0,34723
Медіана	0,351032	0,339717	0,338006
Коефіцієнт кореляції на сукупності	0,354508		

Таблиця 3 – Характеристики варіації коефіцієнта кореляції при слабкій залежності

Оцінюваний параметр	Розмір вибірки		
	32	16	8
Мінімім	-0,34711	-0,61938	-0,86307
Максимум	0,265146	0,391063	0,50781
Середнє	-0,06067	-0,09445	-0,12498

Продовж. табл. 3

Медіана	-0,07199	-0,06099	-0,06883
Коефіцієнт кореляції на сукупності	-0,04736		

Характер зміни розкиду коефіцієнтів кореляції і відповідних параметрів положення для слабкої залежності добре видно на рис. 2. У цілому характер зміни для різної сили кореляційної залежності подібний і тому не приводиться.

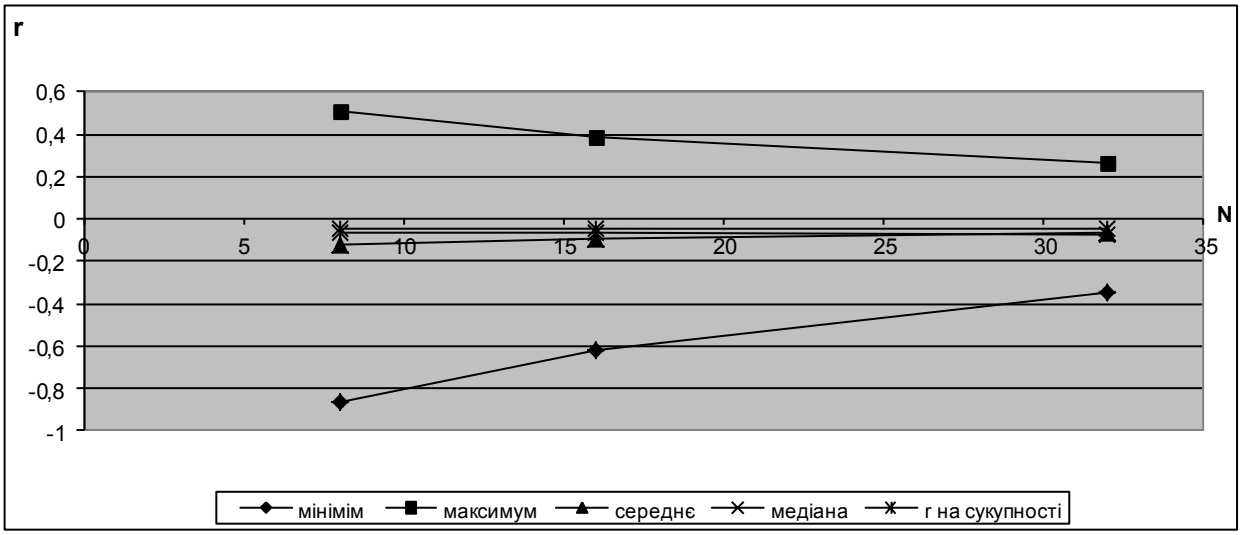

Рисунок 2 – Характеристики розкиду значень коефіцієнта для слабкого зв'язку між факторами

З проведеного обчислювального експерименту можна зробити такі висновки.

При високій тісноті зв'язку зберігається стійкість середнього значення коефіцієнта кореляції незалежно від розміру вибірки. Розсіяння вибіркових значень збільшується при зменшенні розміру вибірки.

При помірній тісноті зв'язку зі зменшенням розміру вибірки збільшується розмах можливих відхилень. Можлива навіть зміна знаку.

Для високої і помірної тісноти зв'язку стійкість середнього значення коефіцієнта кореляції і медіани зберігається незалежно від розміру вибірки.

При слабкому зв'язку стійкість відсутня аж до можливої зміни знаку незалежно від розміру вибірки. Середнє значення сильно змінюється, але перевірка приналежності значень кореляції для вибірок 32 і 8 підтверджує їх належність до однієї сукупності. Медіана зберігає стійкість.

Крім того, з приведених результатів можна зробити висновок, що встановлювати розмір екзаменаційної просто в частках від навчальної (незалежно від її розміру і величини зв'язку регресорів з відгуком) у загальному випадку не дає можливості порівняння значень вибіркових коефіцієнтів кореляції для обґрунтованого формування структури моделі.

Таким чином, при малих розмірах вибірок і слабкому зв'язку окреме значення коефіцієнта кореляції не може бути обґрунтовано прийнято для визначення специфікації.

Виникає питання, чи можливе використання визначення за вибіркою середнього значення коефіцієнта кореляції для оцінки, яка більш обґрунтована для прийняття рішень. Оскільки в реальному експерименті ми маємо тільки одну вибірку, то виникає питання у способах формування деяких підвибірок для визначення середнього чи медіани. Розглянемо два способи формування підвибірок для такої оцінки.

У табл. 4 приведені розрахунки для попередньої задачі методом «складного ножа». При цьому вся вибірка виступає як генеральна сукупність, а перші 32 рядки – як експери-

ментальна вибірка. В табл. 5 характеристики отримані за результатами розрахунків за 50 підвибірок розміром 16 з перших 16 рядків.

Видно, що, по-перше, отримані результати мало відрізняються між собою, а, по-друге, вони ближчі до значення вибірки, а не до генеральної сукупності.

Таблиця 4 – Характеристики визначення коефіцієнта кореляції для «складного ножа»

Номер змінної	1	2	3	4
Коефіцієнт кореляції на сукупності	0,843617	0,354508	-0,04736	-0,06843
Коефіцієнт кореляції на вибірці	0,848508	0,260155	-0,08454	-0,1664
Медіана	0,849543	0,269019	-0,0854	-0,16038
Середнє квадратичне	0,006478	0,024187	0,033173	0,029202
Мінімум	0,835397	0,197461	-0,1758	-0,2272
Максимум	0,861366	0,304945	-0,00187	-0,10628

У табл. 6 представлено розрахунки коефіцієнта кореляції, виконані за методом «складного ножа» для реальної задачі, описаної в [1, 8], з ортогональною матрицею плану експеримента.

Таблиця 5 – Характеристики визначення коефіцієнта кореляції для множини випадкових підвибірок

Номер змінної	1	2	3	4
Коефіцієнт кореляції на сукупності	0,843617	0,354508	-0,04736	-0,06843
Коефіцієнт кореляції на вибірці	0,848508	0,260155	-0,08454	-0,1664
Медіана	0,838435	0,269288	-0,10797	-0,1673
Середнє значення	0,84413	0,263974	-0,06761	-0,18882
Мінімум	0,753475	-0,07872	-0,40104	-0,6503
Максимум	0,928319	0,60517	0,257844	0,265488

Таблиця 6 – Характеристики розрахунків для методу «складного ножа» для реальної задачі

	Номер змінної	Коефіцієнт кореляції	Медіана	Середнє квадратичне	Мінімум	Максимум
X1	1	0,548932	0,574522	0,029453	0,54034	0,62234
Z1	2	-0,00983	-0,0143	0,035844	-0,05665	0,052414
X2	3	0,736251	0,763383	0,031097	0,753704	0,795211
Z2	4	-0,0295	-0,04497	0,033518	-0,08396	0,022383
X3	5	-0,00287	-0,01597	0,045503	-0,08409	0,089527
Z3	6	0,021653	0,023607	0,044325	-0,04383	0,149796
X1X2	7	0,288952	0,294381	0,047677	0,198973	0,389873
X1X3	8	0,000993	0,00306	0,046595	-0,12977	0,149704
X2X2	9	-0,00806	-0,01515	0,061019	-0,15031	0,128829
X1X2X3	10	0,004787	-0,00375	0,057829	-0,17151	0,179035
Z1Z2	11	0,003797	0,010691	0,028866	-0,03299	0,047001
Z1Z3	12	0,011087	0,016728	0,038013	-0,05422	0,092625
Z2Z3	13	-0,00473	9,24E-06	0,032088	-0,06004	0,030103

3. Висновки

1. Виділення за коефіцієнтом кореляції при значенні менше 0,2 сумнівне і необґрунтоване. Оскільки в реальних задачах по побудові регресійних моделей для значної кількості регресорів мають місце саме такі коефіцієнти кореляції, то, відповідно, по-перше, необхідні до-

даткові перевірки обґрунтованості включення таких регресорів у модель, а, по-друге, урахування цього факту при аналізі і використанні моделі. Слід відзначити, що правило виконується, так як процедури вибору окремої структури (конкретна специфікація) зазвичай багатоступінчасті, а також це фактично забезпечується тим, що, як правило, при експоненціальній формі розподілу сили впливу регресорів [2, 10] на «хвіст» моделі припадає незначна частина частки впливу, яка пояснюється моделлю.

2. У тому випадку, коли матриця експерименту є пасивною або випадковою, бажано використовувати для прийняття рішення не значення коефіцієнта, розрахованого за даною вибіркою, а медіану, яка розрахована методом «складного ножа» або методом формування випадкових підвибірок з цієї вибірки.

3. Розмір екзаменуючої вибірки при застосуванні для конкретної специфікації двох вибірок може бути 0,25 від навчальної тільки в тому випадку, коли вона складає не менше 30 експериментів. В інших випадках необхідно приймати не менше, ніж 0,5.

4. При використанні екзаменуючої вибірки для прийняття рішення рекомендується використовувати не значення коефіцієнта, розрахованого за даною вибіркою, а медіану, розраховану методом «складного ножа» або формування випадкових підвибірок з цих вибірок.

СПИСОК ДЖЕРЕЛ

1. Лапач С.М. Проблеми побудови регресійних моделей процесів різання металів. *Вісник НТУУ «КПІ». Машинобудування.* Київ. 2014. № 3 (72). С. 40–47.

2. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. К.: Морион, 2002. 640 с.

3. Ивахненко А.Г., Мюллер Й.А. Самоорганизация прогнозирующих моделей. К.: Техніка, 1984; Берлин: ФЕБ Ферлаб Техник, 1984. 223 с.

4. Іванів П.В., Лапач С.М. Застосування кореляційного аналізу для конкретної специфікації регресійної моделі. *Збірка тез доповідей загальноуніверситетської наук.-техн. конф. молодих вчених та студентів, присвяченої дню науки. Машинобудування. Технологія машинобудування.* К.: НТУУ «КПІ», 2015. С. 75–77.

5. Pardoux C. Sur la selection de variables en regression multiple. *Cah. Bur. Univ. rech. oper*. 1982. N 39–40. P. 101–133.

6. Вентцель Е.С. Теория вероятностей. М.: Высшая школа, 2000. 576 с.

7. Лапач С.Н., Пасечник М.Ф., Чубенко А.В. Статистические методы в фармакологии и маркетинге фармацевтического рынка. К.: ЗАТ «Укрспецмонтаж», 1999. 312 с.

8. Кацев П.Г. Статистические методы исследования режущего инструмента. 2-е изд., перераб. и доп. М.: Машиностроение, 1974. 231 с.

9. Соболь И.М., Статников Р.Б. Выбор оптимальных параметров в задачах со многими критериями. М.: Наука, 1981. 111 с.

10. Satterthwaite F.E. Random Balance Experimentation. *Technometrics*. 1959. Vol. 1, N 2. P. 111–137.

Стаття надійшла до редакції 28.06.2018