Журнал обчислювальної та прикладної математики **2017, №2(125)** Journal of Numerical & Applied Mathematics

УДК 519.21
MSC 62G15

CONSTRUCTION OF CONFIDENCE INTERVALS FOR VARIANCE WITH UNKNOWN MEAN BASED ON THREE-SIGMA RULE

Tupko Natalia[1], Vasil'ev Alexander[2], Tupko Olha[3]

[1]Educational and Scientific Institute of Computer Information Technologies, National Aviation University, Kyiv, Ukraine, E-mail: natupko@ukr.net
[2]Institute of Mathematics, Economics and Mechanics, Odessa National University of I. I. Mechnikov, Odessa, Ukraine, E-mail: av5111955@gmail.com
[3]Char of Mathematical Finance, Technical University of Munich, Munich, Germany, E-mail: olha.tupko@gmail.com

МОДЕЛЮВАННЯ ДОВІРЧИХ ІНТЕРВАЛІВ ДЛЯ ДИСПЕРСІЇ З НЕВІДОМИМ МАТЕМАТИЧНИМ СПОДІВАННЯМ НА ОСНОВІ ПРАВИЛА З СІГМА

Н. П. Тупко, О. Б. Васильєв, О. С. Тупко

[1]Навчально-науковий інститут Комп'ютерних інформаційних технологій, Національний Авіаційний Університет, Київ, Україна, E-mail: natupko@ukr.net
[2]Інститут математики, економіки і механіки, Одеський національний університет імені І. І. Мечникова, Одеса, Україна, E-mail: av5111955@gmail.com
[3]Факультет Фінансової математики, Технічний Університет Мюнхена, Мюнхен, Німеччина, E-mail: olha.tupko@gmail.com

Abstract. Based on variance estimation in case of unknown mean [3], confidence intervals were built using rule 3σ. Using software implementation in R, significance levels for four distributions (Standard Normal, Exponential, Uniform, Poisson) were calculated.
Keywords: Variance, mean, estimation, confidence intervals, mean squared error.

Резюме. Використовуючи оцінки для дисперсії у випадку невідомого математичного сподівання [3], побудовано довірчі інтервали для невідомої дисперсії за допомогою правила 3σ. Підраховано відповідні рівні значущості для деяких розподілів за допомогою програмної реалізації.
Ключові слова: дисперсія, математичне сподівання, оцінка, довірчі інтервали, середній квадрат похибки.

Вступ

Для побудови довірчих меж для основної розподіленої маси значень генеральної сукупності, що відповідають заданим рівням значущості, а також для довірчого оцінювання невідомих параметрів вкрай необхідні математичне сподівання, дисперсія, коефіцієнти коваріації. У свою чергу, довірчі

інтервали або межі є основою для побудови статистичних критеріїв, що засновані на навчаючих вибірках і використовуються у теорії розпізнавання образів або класифікації об'єктів. Такі задачі виникають при диференціальній діагностиці онкологічних захворювань (рак молочної залози або фіброаденоматоз, аденокарценома щитовидної залози або вузловий зоб чи аутотіреоїдит, рак шлунку або непухлинне захворювання шлунку тощо), які постійно перебували у колі наукових інтересів Юрія Івановича Петуніна. На жаль, класичні методи теорії перевірки гіпотез за допомогою статистичних критеріїв, зокрема теорія Неймана-Пірсона, що дає один з найбільш потужніх критеріїв, а також її різні модифікації (оптимальні статистичні критерії, критерії, що використовують процедуру неприйняття рішення, індивідуальні статистичні критерії) базуються на функціях розподілу генеральних сукупностей, які майже ніколи не відомі на практиці. У зв'язку з цим можна використовувати лише ту інформацію, що можливо одержати на основі навчаючих вибірок. Побудова оцінок для функцій розподілу або щільностей імовірностей вимагає навчаючих вибірок великого об'єму, що містять декілька сотень або тисяч вибіркових значень. Це є у переважній більшості випадків задачею, яку неможливо виконати, оскільки одержання та дослідження кожного вибіркового значення пов'язано з певними матеріальними витратами. На відміну від цього, довірчі межі та інтервали, як правило, грунтуються на знанні математичного сподівання, дисперсії або коефіцієнта коваріації, які можна визначити за допомогою малих чи середніх вибірок, що мають лише від кількох десятків до двохсот спостережень. З цього випливає, що проблема оцінки невідомих математичних сподівань, дисперсії коефіцієнта коваріаціїї є дуже важливою та актуальною задачею для теорії розпізнавання образів, яка зараз широко використовується у техніці, біології, медицині, соціології та інших прикладних науках.

Використовуючи оцінки для дисперсії у випадку невідомого математичного сподівання [3], у роботі побудовано довірчі інтервали для невідомої дисперсії за допомогою правила 3σ, обгрунтованого Ю. І. Петуніним і Д. Ф. Височанським у 1979 році [8].

Оцінки невідомої дисперсії у випадку невідомого математичного сподівання

Для дисперсії розглянемо дві оцінки — незміщену

$$\tilde{S}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 \tag{1}$$

і зміщену

$$S_*^2 = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^2 - \bar{x}^2\right), \tag{2}$$

математичне сподівання яких дорівнює

$$m\left(\tilde{S}^2\right) = \sigma^2, \qquad m\left(S_*^2\right) = \frac{n-1}{n}\sigma^2. \tag{3}$$

Середній квадрат похибки незміщеної оцінки $\tilde{S}^2$, а також різниця цих похибок згідно результатам, отриманим у роботі [3], дорівнює

$$\begin{gathered}\delta^2\left(\tilde{S}^2\right)=D\left(\tilde{S}^2\right)=\frac{1}{n}m(x^4)+\frac{(n-1)^2+2}{n(n-1)}\left(m^2+\sigma^2\right)^2+\\+\frac{4-2(n-1)}{n(n-1)}(n-2)m^2\left(m^2+\sigma^2\right)-\frac{4}{n}m(x^3)m+\frac{(n-2)(n-3)}{n(n-1)}m^4-\\-\sigma^4=\frac{1}{n}m(x^4)-\frac{4}{n}m(x^3)m+m^4\frac{3}{n}+m^2\sigma^2\frac{6}{n}+\sigma^4\frac{3-n}{n(n-1)},\end{gathered}\tag{4}$$

$$\begin{gathered}V=D\left(\tilde{S}^2\right)-\delta^2\left(S_*^2\right)=\frac{2n-1}{n^2}\times\\\times\left[m\left(x^4\right)+3m^4+6m^2\sigma^2+\frac{-3n^2+8n-3}{n(n-1)(2n-1)}\sigma^4-4m(x^3)m\right].\end{gathered}\tag{5}$$

Відповідно знаходимо середній квадрат похибки зміщеної оцінки

$$\begin{gathered}\delta^2\left(S_*^2\right)=\frac{(n-1)^2}{n^2}\left[\frac{1}{n}m(x^4)+\frac{(n-1)^2+2}{n(n-1)}(m^2+\sigma^2)^2+\right.\\+\frac{4-2(n-1)}{n(n-1)}(n-2)m^2(m^2+\sigma^2)-\frac{4}{n}m(x^3)m+\\\left.+\frac{(n-2)(n-3)}{n(n-1)}m^4\right]+\frac{2-n}{n}\sigma^4=\\=\frac{(n-1)^2}{n^2}\left[\frac{1}{n}m(x^4)-\frac{4}{n}m(x^3)m+m^4\frac{3}{n}+m^2\sigma^2\frac{6}{n}\right]+\sigma^4\frac{5n-n^2-3}{n^3},\end{gathered}\tag{6}$$

де σ^2 — дисперсія, m — математичне сподівання, $m(x^k)$ — момент k-го порядку.

У роботі [3] доведено, що для нормально та рівномірно розподілених випадкових величин $\delta^2\left(S_*^2\right)<\left(\tilde{S}^2\right)$, тобто зміщена оцінка S_*^2 є більш точною ніж незміщена $\tilde{S}^2$.

Перевіримо це твердження на прикладі інших розподілів:

— для експоненціально розподілених величин з параметром λ. У цьому випадку $m(x)=\frac{1}{\lambda}$, $m(x^2)=\frac{2}{\lambda^2}$, $m(x^3)=\frac{6}{\lambda^3}$, $m(x^4)=\frac{24}{\lambda^4}$, $\sigma^2=\frac{1}{\lambda^2}$. Відповідно

$$V=\frac{1}{\lambda^4}\frac{2n^2(9n-15)+17n-3}{n^3(n-1)}$$

і при $n\geq 2$ $V>0$, тобто зміщена оцінка S_*^2 є більш точною ніж незміщена $\tilde{S}^2$;

— для пуассонівських випадкових величин з параметром λ. У цьому випадку $m(x)=\lambda$, $m(x^2)=\lambda^2+\lambda$, $m(x^3)=\lambda^3+3\lambda^2+\lambda$, $m(x^4)=\lambda^4+6\lambda^3+7\lambda^2+\lambda$, $\sigma^2=\lambda$. Відповідно

$$V=\lambda^2\frac{6n^2(n-2)+11n-3}{n^3(n-1)}+\lambda$$

і при $n \geq 2$, $V > 0$, тобто зміщена оцінка S_*^2 є більш точною ніж незміщена $\tilde{S}^2$.

Побудова довірчих інтервалів для невідомої дисперсії на основі правила 3σ у випадку невідомого математичного сподівання

Якщо випадкова величина x нормально розподілена, то згідно для випадкової величини справедливо правило трьох сігм, тобто

$$P\left\{-3\sigma + x < m(x) < 3\sigma + x\right\} > 0.95, \tag{7}$$

де $\sigma = \sqrt{D(x)}$ — середньо квадратичне відхилення.

Аналогічно побудуємо довірчі інтервали для дисперсії довільного розподілу на базі оцінок S_*^2 та $\tilde{S}^2$ і підрахуємо рівні значущості.

Нехай вибірка $x_1, x_2, \ldots, x_n$ — послідовність значень незалежних однаково розподілених випадкових величин.

Згідно (3) на основі (7) побудуємо наступні співвідношення

$$-3\sigma\left(\tilde{S}^2\right) + \tilde{S}^2 < \sigma^2 < 3\sigma\left(\tilde{S}^2\right) + \tilde{S}^2, \tag{8}$$

$$-\frac{3n}{n-1}\delta\left(S_*^2\right) + \frac{n}{n-1}S_*^2 < \sigma^2 < \frac{3n}{n-1}\delta\left(S_*^2\right) + \frac{n}{n-1}S_*^2. \tag{9}$$

Тоді, враховуючи (4) та (6) і використовуючи у цих формулах замість центрального моменту k-го порядку $L_k = m(x^k)$ оцінку

$$\frac{1}{n}\sum_{i=1}^{n} x_i^k,$$

завжди можна побудувати довірчі інтервали для невідомої дисперсії σ^2 на основі співвідношень (8) або (9)

$$Y_1 = \left(-3\sigma\left(\tilde{S}^2\right) + \tilde{S}^2, 3\sigma\left(\tilde{S}^2\right) + \tilde{S}^2\right), \tag{10}$$

$$Y_2 = \left(-\frac{3n}{n-1}\delta\left(S_*^2\right) + \frac{n}{n-1}S_*^2, \frac{3n}{n-1}\delta\left(S_*^2\right) + \frac{n}{n-1}S_*^2\right), \tag{11}$$

де

$$\delta^2\left(\tilde{S^2}\right) = \frac{1}{n^2}\sum_{i=1}^{n} x_i^4 - \frac{4}{n^3}\sum_{i=1}^{n} x_i^3 \sum_{i=1}^{n} x_i + \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^4 \frac{3}{n} + \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \times$$
$$\times \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right)\frac{6}{n} + \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right)^2 \frac{3-n}{n(n-1)},$$

$$\delta^2\left(S_*^2\right) = \frac{(n-1)^2}{n^2}\left[\frac{1}{n^2}\sum_{i=1}^{n} x_i^4 - \frac{4}{n^3}\sum_{i=1}^{n} x_i^3 \sum_{i=1}^{n} x_i + \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^4 \frac{3}{n} + \right.$$
$$\left. + \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right)\frac{6}{n}\right] +$$
$$+ \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right)^2 \frac{5n - n^2 - 3}{n^3}.$$

Обчислення рівнів значущості довірчих інтервалів для невідомої дисперсії

Обчислимо рівні значущості довірчих інтервалів для певного нормального, рівномірного, експоненціального та пуассонівського розподілів.

Будуємо m довірчих інтервалів (10) на основі вибірок розмірності N конкретного розподілу і підраховуємо рівень значущості як $1 - \frac{k}{m}$, де k — кількість інтервалів, у які попадає дисперсія конкретного розподілу. Аналогічно на основі цих вибірок підраховуємо рівні значущості для довірчих меж (11).

Отримали наступні результати.

1. Нормальний розподіл $N(0, 1)$

m	N	Рівень значущості для Y1	Рівень значущості для Y2
5	10	0	0
5	20	0	0
5	40	0	0
5	70	0	0
10	**10**	**0.2**	**0.1**
10	20	0.2	0.2
30	10	0.0333	0.0333
30	20	0.1333	0.1333
30	40	0.0667	0.0667
50	**10**	**0.18**	**0.16**
50	**20**	**0.04**	**0.02**
100	**10**	**0.12**	**0.11**
100	**20**	**0.07**	**0.06**
100	40	0.02	0.02
100	100	0.02	0.02

2. Експоненціальний розподіл з параметром $\lambda = 1$

m	N	Рівень значущості для Y1	Рівень значущості для Y2
5	10	0.2	0.2
5	20	0.2	0.2
5	40	0.2	0.2
5	70	0	0
10	10	0.3	0.3
10	20	0.4	0.4
30	**10**	**0.2333**	**0.2**
30	20	0.1333	0.1333
30	40	0.0333	0.0333
50	10	0.34	0.34
50	**20**	**0.2**	**0.18**
100	10	0.3	0.3
100	**20**	**0.19**	**0.18**
100	40	0.15	0.15
100	100	0.06	0.06

3. Рівномірний розподіл на інтервалі [0;1]

m	N	Рівень значущості для Y1	Рівень значущості для Y2
5	10	0	0
5	20	0	0
5	40	0	0
5	70	0	0
10	10	0.1	0.1
10	20	0	0
30	10	0	0
30	20	0	0
30	40	0	0
50	**10**	**0.08**	**0.06**
50	20	0	0
100	**10**	**0.04**	**0.03**
100	20	0.05	0.05
100	40	0	0
100	100	0.01	0.01

4. Розподіл Пуассона з параметром $\lambda = 1$

m	N	Рівень значущості для Y1	Рівень значущості для Y2
5	10	0.2	0.2
5	20	0	0
5	40	0	0
5	70	0	0
10	10	0	0
10	20	0.2	0.2
30	10	0.2333	0.2333
30	20	0.2333	0.2333
30	40	0.0333	0.0333
50	10	0.06	0.06
50	**20**	**0.12**	**0.1**
100	10	0.14	0.14
100	**20**	**0.1**	**0.09**
100	40	0.09	0.09
100	100	0.03	0.03

Висновок

Отримані у роботі експериментальні результати підтверджують теоретичні викладки, Отже, для запропонованих розподілів оцінка S_*^2 є більш точною ніж $\tilde{S}^2$.

Література

1. Байдак Г. И., Браверман М. Ш., Петунин Ю. И. Аддитивность дисперсии характеристическое свойство гильбертового пространства // Функциональный анализ и его приложения. — 1983. — 17, вып. 3. — С. 66–68.

2. Петунин Ю. И. Приложение теории случайных процессов в биологии и медицине — К.: Наукова думка, 1981. — 320 с.
3. Петунин Ю. И., Тупко Н. П. Теория квадратичных оценок дисперсии // Український математичний журнал.— 1999.— Том 51, № 9.— С. 1217–1231.
4. Курицын Ю. Г., Петунин Ю. И. К теории линейных оценок математического ожидания случайного процесса // Теория вероятности и математическая статистика. — 1970. — Вып. 3. — С. 80–92.
5. Ван дер Варден Б. Л. Математическая статистика — М.: Изд-во иностр. лит., 1960. — 436 с.
6. Крамер Г. Математические методы статистики — М.: Мир, 1975. — 648 с.
7. Вознесенский В. Статистические решения в технологических задачах — Кишенёв: Картя Молдовеняске, 1969. — 232 с.
8. Высочанский Д. Ф., Петунин Ю. И. Обоснование правила 3-sigma для одномодальных распределений // Теория вероятностей и мат. статистика. — 1979. — Вып. 21. — С. 23–35.

Надійшла 01.09.2017