Збірник праць Ін-ту математики НАН України 2017, т. 14, № 4, 7–18

УДК 517.927

Я. В. Гончаренко, Н. В. Дивляш

Національний педагогічний університет імені М. П. Драгоманова, Київ; ya.v.goncharenko@npu.edu.ua

Алгоритм побудови послідовності значень випадкової величини, що має сингулярний розподіл салемівського типу

In the paper we describe algorithm to obtain a sequence of values of random variable ξ with independent identically distributed binary digits having a singular distribution of Salem type. Some examples of application of this algorithm are also given. We consider distribution of ξ as one-parameter distribution that depends on parameter $\rho = \mathbb{P}\{\alpha_i = 1\}$, where α_i is a ith binary digit of ξ. For samples obtained by simulation modeling, we demonstrate three methods for estimation of unknown parameter ρ: method of moments, modified minimum distance estimation and maximum-likelihood estimation methods. We also compare accuracy of obtained estimators.

Key words: singular probability distribution, statistical estimator, simulation modeling.

У роботі наведені опис і приклади застосування алгоритму отримання послідовності значень випадкової величини ξ з незалежними однаково розподіленими двійковими цифрами, що має сингулярний розподіл салемівського типу. При цьому розподіл ξ розглядається як однопараметричний розподіл, що залежить від параметра $\rho = \mathbb{P}\{\alpha_i = 1\}$, де α_i — i-та двійкова цифра ξ. Для отриманих за допомогою імітаційного моделювання вибірок продемонстровано три методи оцінювання невідомого параметра ρ: метод моментів і модифікації методів мінімальної відстані та максимальної правдоподібності, і здійснено порівняння точності отриманих оцінок.

Ключові слова: сингулярний розподіл, статистична оцінка, імітаційне моделювання.

Вступ

На сьогодні отримано ряд результатів, що стосуються дослідження сингулярних розподілів, описано їхні класи, здійснено класифікацію за трьома чистими класами, отримано умови належності до кожного класу [1, 4]. Проте дослідження статистичних задач, пов'язаних із сингулярними розподілами, нині практично відсутні.

У роботах [2, 3] для найпростішого сингулярного розподілу салемівського типу було обґрунтовано модифікації статистичних методів отримання оцінок, а також доведено існування й побудована реалізація оптимального критерію для перевірки простої гіпотези щодо значення параметра розподілу в класі досліджуваних розподілів, на основі якої сформульовані алгоритми статистичних гіпотез щодо параметра розподілу. Ці задачі були розв'язані для розподілу випадкової величини η з незалежними однаково розподіленими двійковими цифрами η_k, $P\{\eta_k = 1\} = p_1$, $P\{\eta_k = 0\} = p_0$, $0 \leq p_i \leq 1$, $p_0 + p_1 = 1$. Відомо [1], що при $p_0 \neq p_1$ випадкова величина η має сингулярний розподіл салемівського типу. При цьому її функція розподілу має вигляд:

$$F_\eta(x) = \beta_{\alpha_1} + \sum_{k=1}^{\infty} \beta_{\alpha_{k+1}} \prod_{j=1}^{k-1} p_{\alpha_j}, \qquad (1)$$

де α_k — k-та двійкова цифра x; $\beta_0 = 0$, $\beta_1 = 1 - p_1$. Якщо як параметр розглянути $p_1 = \rho$, то функцію розподілу (1) можна записати у вигляді

$$F_\eta(x) = \beta_{\alpha_1} + \sum_{k=1}^{\infty} \beta_{\alpha_{k+1}} \rho^{\sum\limits_{j=1}^{k} \alpha_j} (1-\rho)^{k - \sum\limits_{j=1}^{k} \alpha_j}. \qquad (2)$$

Для класу параметричного розподілу (2) було побудовано методи оцінки значення параметра ρ: модифікації методу мінімальної відстані й максимальної правдоподібності, досліджено властивості отриманих оцінок, зокрема для оцінки максимальної правдоподібності доведено її незміщенність, консистентність і ефективність.

Зазначимо, що при побудові прикладів, що ілюструють застосування запропонованих у [2] методів оцінок невідомого параметра однопараметричного сингулярного розподілу (2), виникає проблема: отримати (згенерувати) вибірку, що має досліджуваний сингулярний розподіл, вирішенню якої присвячена ця стаття.

1. Методи оцінювання невідомих параметрів сингулярного розподілу салемівського типу

Нагадаємо, у чому полягає задача оцінки параметра розподілу. Нехай ми маємо реалізацію $X_n = (x_1, x_2, \dots, x_n)$ вибірки з розподілом $F(\cdot; \theta)$. Розподіл залежить від деякого параметра θ з множини всіляких значень Θ, його значення невідоме і його потрібно оцінити (визначити) за реалізацією X_n вибірки.

Для оцінки параметра θ у просторі X — множині реалізацій вибірки — необхідно визначити (побудувати) функцію $h(\cdot)$ зі значеннями в Θ — множині всеможливих значень параметра θ, таку, що значення $h(X_n)$ дорівнює θ (або хоча б наближено дорівнює θ). Значення $\hat{\theta} = h(X_n)$ ми й використовуватимемо як невідоме θ. Причому варто зазначити, що для кожної реалізації X_n значення $\hat{\theta} = h(X_n)$, яке використовується як θ, своє (функція $\hat{\theta}$ є випадковою величиною).

Для оцінювання потрібно врахувати, наскільки велика похибка $\hat{\theta} - \theta$, тобто наскільки сильно можуть відрізнятися значення оцінки $\hat{\theta} = h(X_n)$ від оцінюваної величини θ.

Оцінка $\hat{\theta}$ називається *незміщеною оцінкою параметра* θ, якщо $M(\hat{\theta}) = \theta$, або, що те ж саме, $M(\hat{\theta} - \theta) = 0$.

Послідовність оцінок $\{\hat{\theta_n}\}$ називається *консистентною послідовністю оцінок параметра* θ, якщо для кожного $\varepsilon > 0$:

$$P\{|\hat{\theta_n} - \theta| < \varepsilon\} \to 1(n \to \infty).$$

Незміщеною оцінкою θ^* параметра θ називається його *найкраща оцінка, оцінка мінімальної дисперсії або ефективна оцінка*, якщо $D\theta^* = \inf\limits_{\hat{\theta}: M\hat{\theta} = \theta} D\hat{\theta}$.

Зупинимося детальніше на алгоритмах модифікованих методів мінімальної відстані й максимальної правдоподібності.

1.1. Модифікація методу мінімальної відстані

1. Розбити відрізок $[0;1]$ на 2^m двійкових відрізків рангу m:

$$\Delta_{\alpha_1\alpha_2\ldots\alpha_m} = \left[\sum_{i=1}^{m} \alpha_i 2^{-i}; \sum_{i=1}^{m} \alpha_i 2^{-i} + 2^{-m}\right], \quad \alpha_i \in \{0;1\}.$$

2. Знайти відносні частоти $\nu(\Delta_{\alpha_1\alpha_2\ldots\alpha_m})$ попадання вибіркових значень x_i, $i = \overline{1,n}$, у двійкові відрізки рангу m.

3. Знайти значення ймовірностей для кожного з двійкових відрізків:

$$P\{x \in \Delta_{\alpha_1\alpha_2\ldots\alpha_m}\} = \varrho^{\sum\limits_{j=1}^{m}\alpha_j}(1-\varrho)^{m-\sum\limits_{j=1}^{m}\alpha_j}.$$

4. Для всіх $\Delta_{\alpha_1\alpha_2\ldots\alpha_m}$ знайти найбільші значення функцій

$$|\varphi_{\Delta_{\alpha_1\alpha_2\ldots\alpha_m}}(\varrho)| = \left|\varrho^{\sum\limits_{j=1}^{m}\alpha_j}(1-\varrho)^{m-\sum\limits_{j=1}^{m}\alpha_j} - \nu(\Delta_{\alpha_1\alpha_2\ldots\alpha_m})\right|, \quad \varrho \in [0;1].$$

5. Знайти двійковий відрізок $\Delta^*_{\alpha_1\alpha_2\ldots\alpha_m}$ рангу m, для якого $\max\limits_{\Delta_{\alpha_1\alpha_2\ldots\alpha_m}} |\varphi_{\Delta_{\alpha_1\alpha_2\ldots\alpha_m}}(\varrho)|$ набуває найменшого значення при $\varrho \in (0;1)$.

6. На $\Delta^*_{\alpha_1\alpha_2\ldots\alpha_m}$ обчислити значення оцінки

$$\widehat{\varrho} = \frac{\sum\limits_{i=1}^{n}\alpha_j}{m}. \tag{3}$$

Зауваження 1. *Метод мінімальної відстані може використовуватися при достатньо великих m і дає незміщену оцінку при $m \to \infty$.*

1.2. Модифікація методу максимальної правдоподібності

1. Розбити відрізок $[0;1]$ на 2^m двійкових відрізків рангу m:

$$\Delta_{\alpha_1\alpha_2\ldots\alpha_m} = \left[\sum_{i=1}^{m}\alpha_i 2^{-i}; \sum_{i=1}^{m}\alpha_i 2^{-i} + 2^{-m}\right], \quad \alpha_i \in \{0;1\}.$$

2. Знайти двійковий розклад кожного елемента вибірки x_i до m-ї двійкової цифри включно, тобто для будь-якого $i = \overline{1,n}$ знайти послідовність (α_{ij}), $j = \overline{1,m}$:

$$x_i = \frac{\alpha_{i1}}{2} + \frac{\alpha_{i2}}{2^2} + \ldots + \frac{\alpha_{in}}{2^n} + \ldots + \frac{\alpha_{im}}{2^m}.$$

3. Обчислити значення оцінки:

$$\widehat{\varrho} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\alpha_{ij}}{mn}. \qquad (4)$$

У статті [2] було доведено, що незміщеними і консистентними оцінками параметра ρ розподілу (1) є:

1. Вибірковий момент першого порядку:

$$\widehat{\varrho} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (5)$$

2. Оцінка (3), отримана за допомогою модифікації методу мінімальної відстані.

3. Оцінка (4), отримана за допомогою модифікації методу максимальної правдоподібності.

Для оцінки (4) була також доведена її ефективність.

2. Моделювання значень випадкової величини, що має сингулярний розподіл

Найзагальнішим способом отримання послідовності випадкових чисел, що є послідовністю реалізацій випадкової величини, яка розподілена за довільним законом, є спосіб, оснований на процесі формування їх із вихідної послідовності псевдовипадкових чисел. Псевдовипадкові числа — це числа, отримані за деяким правилом (формулою), що імітує значення випадкової величини, яка має рівномірний на $[0, 1]$ розподіл. На сьогодні розроблено низку алгоритмів для отримання псевдовипадкових чисел, зокрема датчики псевдовипадкових чисел, які входять до складу більшості програмних комплексів.

Для перетворення послідовності випадкових чисел, що є реалізаціями випадкової величини з рівномірним законом розподілу на проміжку $[0;1]$, у послідовність випадкових чисел, що є реалізаціями випадкової величини із заданою функцією розподілу $F(x)$, треба із сукупності випадкових чисел із рівномірним законом розподілу обрати випадкове число ξ і розв'язати відносно x рівняння: $F(x) = \xi$.

У нашому випадку необхідно знайти розв'язок рівняння

$$\beta_{\alpha_1} + \sum_{k=1}^{\infty} \beta_{\alpha_{k+1}} \rho^{\sum\limits_{j=1}^{k} \alpha_j} (1-\rho)^{k-\sum\limits_{j=1}^{k} \alpha_j} = \xi. \tag{6}$$

Припустимо, що ρ набуває деякого фіксованого значення $\rho_0 \in (0;1)$. Представимо випадкове число ξ за допомогою його Q-зображення [1] з матрицею $Q = (q_0, q_1)$, де $q_0 = 1 - \rho_0$, $q_1 = \rho_0$.

Тобто розкладемо ξ у ряд виду:

$$\xi = \beta_{\gamma_1} + \sum_{k=2}^{\infty} \beta_{\gamma_{k+1}} \prod_{j=1}^{k-1} q_{\gamma_j}, \tag{7}$$

де $\gamma_k \in \{0, 1\}$, $\beta_0 = 0$, $\beta_1 = q_0 = 1 - \rho_0$.

Вираз (7) можна переписати у формі:

$$\xi = \beta_{\gamma_1} + \sum_{k=1}^{\infty} \beta_{\gamma_{k+1}} \rho_0^{\sum\limits_{j=1}^{k} \gamma_j} (1-\rho_0)^{k-\sum\limits_{j=1}^{k} \gamma_j}. \tag{8}$$

Виникає питання: як, маючи довільне фіксоване значення $\xi \in [0, 1]$, знайти послідовність γ_k його цифр у Q-розкладі?

Для цього можна використоти процедуру, алгоритм якої представлено на рис. 1.

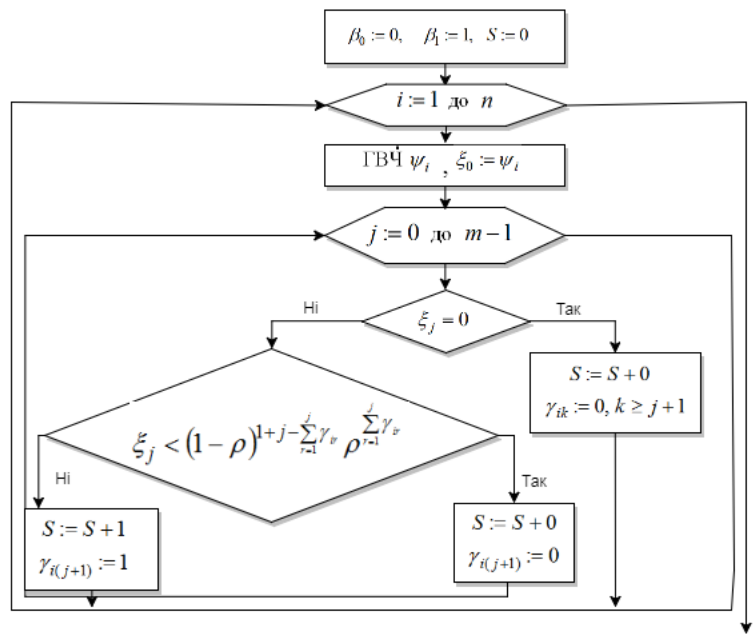

Рис. 1

Нехай $\beta_0 = 0$ і $\beta_1 = 1 - \rho_0$. Якщо справедлива нерівність $\xi < 1 - \rho_0$, то перша цифра $\gamma_1 = 0$, інакше $\gamma_1 = 1$. Позначимо $\xi_1 = \xi - \beta_{\gamma_1}$. Якщо $\xi_1 = 0$, то вважатимемо $\gamma_j = 0$ для всіх $j \geq 2$. Якщо $0 < \xi_1 < (1-\rho_0)(1-\rho_0)^{1-\gamma_1}\rho_0^{\gamma_1}$, то $\gamma_2 = 0$, інакше $\gamma_2 = 1$.

Позначимо $\xi_2 = \xi_1 - \beta_{\gamma_2}(1-\rho_0)^{1-\gamma_1}\rho_0^{\gamma_1}$. Якщо $\xi_2 = 0$, то вважатимемо $\gamma_j = 0$ для всіх $j \geq 3$. Якщо

$$0 < \xi_2 < (1-\rho_0)(1-\rho_0)^{2-(\gamma_1+\gamma_2)}\rho_0^{\gamma_1+\gamma_2},$$

то $\gamma_3 = 0$, інакше $\gamma_3 = 1$.

Продовжуючи подібні міркування, на k-тому кроці запишемо

$$\xi_k = \xi_{k-1} - \beta_{\gamma_k}(1-\rho_0)^{k-1-\sum\limits_{j=1}^{k-1}\gamma_j}\rho_0^{\sum\limits_{j=1}^{k-1}\gamma_j}.$$

Якщо $\xi_k = 0$, то кожна наступна цифра γ_j, починаючи з $(k+1)$-ї, дорівнюватиме нулю. Якщо

$$0 < \xi_k < (1-\rho_0)(1-\rho_0)^{k-\sum\limits_{j=1}^{k}\gamma_j} \rho_0^{\sum\limits_{j=1}^{k}\gamma_j},$$

то k-та цифра $\gamma_k = 0$, інакше k-та цифра $\gamma_k = 1$.

Таким чином, рівність (6) запишемо як:

$$\beta_{\alpha_1}+\sum_{k=1}^{\infty}\beta_{\alpha_{k+1}}\rho_0^{\sum\limits_{j=1}^{k}\alpha_j}(1-\rho_0)^{k-\sum\limits_{j=1}^{k}\alpha_j}=\beta_{\gamma_1}+\sum_{k=1}^{\infty}\beta_{\gamma_{k+1}}\rho_0^{\sum\limits_{j=1}^{k}\gamma_j}(1-\rho_0)^{k-\sum\limits_{j=1}^{k}\gamma_j}. \quad (9)$$

Очевидно, що рівняння (9) перетворюється на правильну рівність за умови $\alpha_i = \gamma_i$ для будь-якого $i = 1, 2, ...$

3. Приклад

За даним алгоритмом ми згенерували вибірку обсягом $n = 250$ із розподілу виду (2).

Для вибіркових значень x_i, $i = \overline{1,250}$ були знайдені всі можливі двійкові відрізки 5-го рангу виду $\Delta_{\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5}$, $\alpha_i \in \{0;1\}$ (дивись табл. 1).

Відносні частоти $\nu_k = \nu(\Delta_{\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5})$, $k = \overline{1,32}$ визначались як відносна частота потрапляння вибіркового значення x_i, $i = \overline{1,250}$ у кожен із циліндрів $\Delta_{\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5}$.

Розвя'яжемо задачу оцінки значення невідомого параметра ρ трьома різними методами.

1. Метод моментів.

$$\widehat{\varrho_1} = \sum_{k=1}^{32} x_k \nu_k,$$

   де $x_k = \sum\limits_{j=1}^{5} \alpha_{kj} 2^{-j} + \frac{1}{2^6}$, α_{kj} – j-та двійкова цифра x_k.

   Отримаємо $\widehat{\varrho_1} = 0,623875$.

2. Метод мінімальної відстані.

   Запишемо відхилення ймовірностей попадання вибіркових значень x_i, $i=\overline{1,250}$ у двійкові відрізки $\Delta_{\alpha_1\alpha_2\ldots\alpha_5}$ від емпіричних частот $\nu(\Delta_{\alpha_1\alpha_2\ldots\alpha_5})$(дивись табл.1):

$$\varphi(\varrho)=\varrho^{\sum\limits_{j=1}^{m}\alpha_j}(1-\varrho)^{m-\sum\limits_{j=1}^{m}\alpha_j}-\nu(\Delta_{\alpha_1\alpha_2\ldots\alpha_m}).$$

   Найменше значення

$$\min\{\max_{\Delta_{\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5}}|\varphi(\varrho)|\}=0,002332275,$$

   яке досягається на відрізку Δ_{01011}, є значенням функції $|\varphi(\varrho)|$ в точці $\varrho=0,6$.

   Таким чином, ми отримали оцінку за допомогою методу мінімальної відстані

$$\widehat{\varrho}_2=0,6.$$

3. Метод максимальної правдоподібності.

   Використавши формулу (4), отримаємо оцінку

$$\widehat{\varrho}_3=0,6264.$$

Для порівняння *точності отриманих оцінок* застосуємо середньоквадратичні оцінки відхилень теоретичних значень від емпіричних, тобто розрахуємо величини:

$$\sigma_e^2(\widehat{\varrho}_s)=\sum_{\Delta_{\alpha_1\alpha_2\ldots\alpha_m}}(P_{\widehat{\varrho}_s}\{x\in\Delta_{\alpha_1\alpha_2\ldots\alpha_m}\}-P_{\varrho}\{x\in\Delta_{\alpha_1\alpha_2\ldots\alpha_m}\})^2,$$

$s=1,2,3$.

Ми маємо $\sigma_1=0,00612$, $\sigma_2=0,02882$, $\sigma_3=0,00361$, тобто найточнішою є оцінка, отримана за допомогою методу максимальної правдоподібності.

Табл. 1: Обчислення оцінки параметра за методом мінімальної відстані

$\Delta_{\alpha_1\ldots\alpha_5}$	ν_k	$\varphi(\varrho)$	$F(x,\widehat{\varrho}_1)$	$F(x,\widehat{\varrho}_2)$	$F(x,\widehat{\varrho}_3)$
Δ_{00000}	0,008	$(1-\varrho)^5-0,008$	0	0	0
Δ_{00001}	0,012	$\varrho(1-\varrho)^4-0,012$	0,00753	0,01024	0,00728
Δ_{00010}	0,008	$\varrho(1-\varrho)^4-0,008$	0,02001	0,02560	0,01948
Δ_{00011}	0,020	$\varrho^2(1-\varrho)^3-0,020$	0,03250	0,04096	0,03169
Δ_{00100}	0,012	$\varrho(1-\varrho)^4-0,012$	0,05321	0,06400	0,05215
Δ_{00101}	0,020	$\varrho^2(1-\varrho)^3-0,020$	0,06570	0,07936	0,06435
Δ_{00110}	0,024	$\varrho^2(1-\varrho)^3-0,024$	0,08641	0,10240	0,08481
Δ_{00111}	0,036	$\varrho^3(1-\varrho)^2-0,036$	0,10712	0,12544	0,10527
Δ_{01000}	0,012	$\varrho(1-\varrho)^4-0,012$	0,14147	0,16000	0,13958
Δ_{01001}	0,020	$\varrho^2(1-\varrho)^3-0,020$	0,15396	0,17536	0,15178
Δ_{01010}	0,024	$\varrho^2(1-\varrho)^3-0,024$	0,17467	0,19840	0,17224
Δ_{01011}	0,032	$\varrho^3(1-\varrho)^2-0,032$	0,19538	0,22144	0,19270
Δ_{01100}	0,020	$\varrho^2(1-\varrho)^3-0,020$	0,22973	0,25600	0,22701
Δ_{01101}	0,032	$\varrho^3(1-\varrho)^2-0,032$	0,25044	0,27904	0,24747
Δ_{01110}	0,036	$\varrho^3(1-\varrho)^2-0,036$	0,28479	0,31360	0,28177
Δ_{01111}	0,056	$\varrho^4(1-\varrho)-0,056$	0,31914	0,34816	0,31608
Δ_{10000}	0,012	$\varrho(1-\varrho)^4-0,012$	0,37612	0,40000	0,37360
Δ_{10001}	0,020	$\varrho^2(1-\varrho)^3-0,020$	0,38861	0,41536	0,38580
Δ_{10010}	0,024	$\varrho^2(1-\varrho)^3-0,024$	0,40932	0,43840	0,40626
Δ_{10011}	0,028	$\varrho^3(1-\varrho)^2-0,028$	0,43003	0,46144	0,42673
Δ_{10100}	0,020	$\varrho^2(1-\varrho)^3-0,020$	0,46438	0,49600	0,46103
Δ_{10101}	0,036	$\varrho^3(1-\varrho)^2-0,036$	0,48510	0,51904	0,48149
Δ_{10110}	0,032	$\varrho^3(1-\varrho)^2-0,032$	0,51945	0,55360	0,51580
Δ_{10111}	0,060	$\varrho^4(1-\varrho)-0,060$	0,55380	0,58816	0,55010
Δ_{11000}	0,024	$\varrho^2(1-\varrho)^3-0,024$	0,61078	0,64000	0,60762
Δ_{11001}	0,032	$\varrho^3(1-\varrho)^2-0,032$	0,63149	0,66304	0,62808
Δ_{11010}	0,036	$\varrho^3(1-\varrho)^2-0,036$	0,66584	0,69760	0,66239
Δ_{11011}	0,060	$\varrho^4(1-\varrho)-0,060$	0,70020	0,73216	0,69670
Δ_{11100}	0,032	$\varrho^3(1-\varrho)^2-0,032$	0,75718	0,78400	0,75422
Δ_{11101}	0,056	$\varrho^4(1-\varrho)-0,056$	0,79153	0,81856	0,78852
Δ_{11110}	0,060	$\varrho^4(1-\varrho)-0,060$	0,84851	0,87040	0,84604
Δ_{11111}	0,096	$\varrho^5-0,096$	0,096	0,92224	0,90356

Література

[1] *Працьовитий М. В.* Фрактальний підхід у дослідженнях сингулярних розподілів. — Київ: Вид-во НПУ імені М. П. Драгоманова. – 1998. – 296 с.

[2] *Гончаренко Я. В., Дивляш Н. В.* Статистичні методи оцінки параметра для одного однопараметричного класусингулярних розподілів салемівського типу. // Науковий часопис НПУ імені М. П. Драгоманова. Серія 1. Фізико-математичні науки. — К.: Вид-во НПУ ім. М. П. Драгоманова. – 2014, 16(1). – С. 81-94.

[3] *Гончаренко Я. В., Дивляш Н. В.* Перевірка статистичної гіпотези про параметр розподілу салемівського типу. // Науковий часопис НПУ імені М. П. Драгоманова. Серія 1. Фізико-математичні науки. – К.: Вид-во НПУ ім. М. П. Драгоманова. – 2014, 16(2). – С. 100-108.

[4] *Турбин А.Ф., Працевитый Н.В.* Фрактальные множества, функции, распределения. – Киев: Наук. думка. – 1992. – 208 с.