УДК 004.02:519.2 **doi:** https://doi.org/10.20998/2522-9052.2023.2.05

С. В. Гадецька[1], В. Ю. Дубницький[2], Ю. І. Кушнерук[3], О. І. Ходирєв[2]

[1] Харківський національний автомобільно-дорожній університет, Харків, Україна
[2] ННІ "Каразінський банківський інститут" ХНУ імені В. Н. Каразіна, Харків, Україна
[3] Харківський національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна

EXCEL-ОРІЄНТОВАНИЙ КАЛЬКУЛЯТОР ДЛЯ ОБЧИСЛЕННЯ РЕЗУЛЬТАТІВ ЕНТРОПІЙНОГО АНАЛІЗУ ДАНИХ, ЩО РОЗПОДІЛЕНІ ПО КАТЕГОРІЯХ

Анотація. Мета роботи. Розробка EXCEL-орієнтованого калькулятора для обчислення результатів ентропійного аналізу даних, які розподілені по категоріях. **Предмет дослідження –** гістограми довільних законів розподілу та таблиці спряженості 2×2. **Методи дослідження:** Ентропійний та інформаційний аналіз гістограм довільних законів розподілу та таблиць спряженості 2×2. **Отримані результати**. Запропоновано використовувати методи ентропійного аналізу для аналізу даних, що розподілені по категоріях та наведено відомості про структуру excel-орієнтованого калькулятора, призначеного для виконання таких розрахунків. Калькулятор дає можливість обчислювати ентропійні характеристики гістограм, виконувати попарне порівняння ентропій гістограм, визначати відстань між гістограмами, обчислювати інформаційний коефіцієнт кореляції, порівнювати розбіжності між гістограмами. Для таблиць спряженості 2×2 калькулятор дає можливість оцінювати значущість взаємодії фактору рядків та фактору стовпців. Калькулятор визначає значення умовних ентропій для таблиць спряженості 2×2. Запропонований калькулятор у деякій мірі заповнює прогалини в існуючих програмних продуктах та може бутий використаний для обробки методами ентропійного аналізу даних, що розподілені по категоріях.

Ключові слова: ентропія; ентропійний аналіз; інформаційний коефіцієнт кореляції; відстань Хелінгера; відстань Кулльбака-Леблера.

Вступ

Припустимо, що кожен елемент множини $X=\{x_i\}$; $i=1, 2, \ldots, N$ має властивість S_s за умови, що $S_S \in S$ та $|S| \le |X|$. Ці властивості можуть бути визначені в одній з наступних шкал вимірювань:

1) номінальній або шкалі найменувань;
2) порядковій, або шкалі рангів;
3) інтервальній, або шкалі інтервалів.

Властивості, які визначені в шкалі рівних відносин, в даній роботі не розглядаються. Алгебраїчні властивості цих шкал і переважні області застосування розглянуто в [1]. Також розрізняють шкали на метричні та неметричні. Поділ шкал вимірювань на метричні та неметричні залежить від типу досліджуваного емпіричного об'єкта, зокрема, від властивостей вимірюваної величини. Метричні шкали – це шкали, у яких ознаки мають одиниці вимірювання. Неметричні шкали – це шкали, у яких немає одиниць вимірювань, наприклад, якісні ознаки. У загальному випадку результати вимірювань, які виконані з використанням цих шкал, прийнято називати даними, що розподілені по категоріях. В [2] наведено для таких даних наступне визначення: категоризовані дані – це дані, які представлено у вигляді частот спостережень, що потрапили в деякі категорії або класи. Для даних, які визначено в інтервальній шкалі, це визначення співпадає з визначенням гістограми [3, 4] і визначенням групованих або частково групованих вибірок [5]. У тому випадку, коли на категорії розподілено можливі значення двовимірної випадкової величини, то отриману гістограму називають таблицею спряженості $r \times c$.

Прийнято, що r визначає кількість рядків, c – стовпчиків. Відповідно до [3, 4] інтервальним статистичним розподілом можливих значень одновимірної випадкової величини (гістограмою) називатимемо дані, які представлено у вигляді табл. 1.

У цій таблиці прийнято, що нижні індекси L і R визначають ліву та праву межі інтервалів, $n_1, n_{2,\ldots}, n_s$ – кількість спостережень у відповідному інтервалі за умови, що загальна кількість спостережень дорівнює N.

У тому випадку, коли індивідуальні значення x_i невідомі, але відома тільки кількість елементів у кожному інтервалі, які вказано в табл. 1, такий спосіб формування гістограми в [5] названий частковим групуванням. В [6, 7] показано, що при великій (понад пятисот) кількості спостережень необхідно застосовувати методи обробки даних, що використовують поняття теорії інформації. Аналогічні висновки зроблені в [8]. В [9] відмічено, що теоретико-інформаційний аналіз експериментальних даних зарекомендував себе як ефективний інструмент дослідження широкого класу процесів і систем. Мірою кількості інформації в цьому підході служить зміна ентропії даних залежно від досліджуваних чинників. Сукупність таких методів отримала назву ентропійного аналізу. Оскільки в [10] було показано, що результати цього аналізу залежать від кількості інтервалів в гістограмі, то в рамках даної роботи порівнюватись будуть лише гістограми, що містять однакову кількість інтервалів.

В практиці медичних та біологічних досліджень набули широкого поширення таблиці спряженості $r \times c$. Їх загальний вигляд наведено в табл. 2.

Таблиця 1 – **Структура гістограми можливих значень одновимірної випадкової величини**

h	$x_{1L} \le x < x_{1R}$	$(x_{2L} = x_{1R}) < x < x_{2R}$	…	$(x_{SL} = x_{(s-1)R}) < x \le x_{sR}$
N	n_1	n_2	…	n_s

Таблиця 2 – **Структура таблиці спряженості $r \times c$**

X	Y						
	y_1	y_2	…	y_j	…	y_c	
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1c}	$n_{1\bullet} = \sum_{j=1}^{c} n_{1j}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2c}	$n_{2\bullet} = \sum_{j=1}^{c} n_{2j}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ic}	$n_{i\bullet} = \sum_{j=1}^{c} n_{ij}$
…	…	…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rc}	$n_{r\bullet} = \sum_{j=1}^{c} n_{rj}$
	$n_{\bullet 1} = \sum_{i=1}^{r} n_{i1}$	$n_{\bullet 2} = \sum_{i=1}^{r} n_{i2}$	…	$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$	…	$n_{\bullet c} = \sum_{i=1}^{r} n_{ic}$	N
$N = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j}.$							

У рамках даної роботи розглянемо таблиці спряженості 2×2**.** Раніше їх використання для оцінювання ефективності деяких видів фармакологічних досліджень розглянуто в [26].

Таблиця 3 – **Структура таблиці спряженості 2×2**

a	b	$\alpha = a + b = n_{1\bullet}$
c	d	$\beta = c + d = n_{2\bullet}$
$\gamma = a + c = n_{\bullet 1}$	$\delta = b + d = n_{\bullet 2}$	N

Слід зауважити, що символи в комірках табл. 2 та табл. 3 позначають кількість даних, які в них потрапили без вказівок їх чисельних значень.

Різні способи обробки категоризованих даних включені до складу багатьох програмних продуктів, наприклад, AtteStat, Statgraphics, SociometryPro 2.3, TextusPro 1.0, TextAnalyst 2.01, WordStat 1.1, ContentAnalyzer 0.52., OCA, Vortex 10.7, IBM SPSS Statistics 26, Stadia 8.0, Statistica 13.3, StatPlus 5.0, DA-System 5.0, X7.2009. Слід зауважити, що лише в системі AtteStat використовуються деякі методи ентропійного аналізу гістограм, причому тільки для одновимірного закону розподілу. Тому розробку програмного забезпечення для ентропійного аналізу таблиць спряженості вигляду $r \times c$ і виконаного у вигляді EXCEL-орієнтованого калькулятора можна, на нашу думку, вважати актуальним завданням.

Аналіз літератури. Необхідні теоретичні відомості про аналіз таблиць спряженості наведено в роботах [2, 11]. Детально відомості про чисельні методи цього аналізу викладено в [11, 12]. В роботі [13] розглянуто способи об'єднання ентропійного аналізу і аналізу таблиць спряженості. На думку автора цієї роботи найбільш ефективним є поєднання методів ентропійного і логлінійного аналізу таблиць спряженості. Основні принципи логлінійного аналізу описано в роботі [14]. Логлінейний аналіз – це, в основному, метод дослідження багатовимірних таблиць спряженості, тобто таких таблиць, що мають вигляд $r \times c \times k$. Цей метод дозволяє перевірити статистичну значущість різних чинників, присутніх в таблиці спряженості і їх взаємодій. Докладно зміст алгоритмів і чисельні приклади їх застосування описано в [15], їх програмна реалізація доступна в системах SPSS та STATISTICA.

Основна статистична гіпотеза, яку перевіряють при аналізі таблиць спряженості – це гіпотеза про відсутність взаємодії між фактором рядків і фактором стовпців. Якщо ця взаємодія відсутня, то початкові дані повинні бути розподілені рівномірно. Для перевірки цієї гіпотези найчастіше обчислюють величину хі-квадрат із використанням співвідношення:

$$\chi^2_{\phi} = \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}, \qquad (1)$$

де E_{ij} – очікувана кількість (частота) об'єктів, що мають ознаки i та j одночасно в припущенні справедливості рівномірного розподілу:

$$E_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}. \qquad (2)$$

Якщо це припущення, що приймають як гіпотезу H_0 справедливо, то величина хі-квадрат має $\nu = (r-1)(c-1)$ ступеней свободи. В цьому випадку правила для прийняття гіпотез мають вигляд:

$$T_1 = \begin{cases} H_0, & якщо \quad \chi^2_{\phi} < \chi^2_{\alpha,\nu}; \\ H_1, & якщо \quad \chi^2_{\phi} > \chi^2_{\alpha,\nu} \end{cases} \qquad (3)$$

або

$$T_2 = \begin{cases} H_0, & якщо \quad Pv_{\phi} > Pv_{\alpha}; \\ H_1, & якщо \quad Pv_{\phi}. < Pv_{\alpha}, \end{cases} \qquad (4)$$

де α – прийнятий рівень довірчої ймовірнрсті, Pv_{ϕ} – фактично отримана довірча ймовірність.

Починати аналіз таблиць спряженості рекомендовано з використання діагностики Сімонова-Цая. Для цього обчислюють величину:

$$SC = \frac{\left(\chi^2(\nu,\alpha)\right)^{1/2}}{3\left(E^2\right)^{3/2}} \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left|\left(n_{ij} - E_{ij}\right)\right|^3}{E_{ij}^2}. \qquad (5)$$

В умові (5) прийнято, що $\chi^2(v,\alpha)$ – значення оберненої функції розподілу χ^2 з v ступенями свободи і рівнем довірчої ймовірності α. Якщо величина SC > 0,25, то використання критеріїв, заснованих на застосуванні величини χ^2, не рекомендовано. У подальшому аналізі використовують декілька критеріїв.

Критерій відношення правдоподібності є таким:

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \ln\frac{n_{ij}}{E_{ij}}. \qquad (6)$$

При застосуванні цього критерію прийнято, що у випадку коли $n_{ij} = 0$, тобто спостереження з даним поєднанням чинників відсутні, то величину критерію для цієї комірки таблиці спряженості приймають рівною нулю. Величина G^2 розподілена відповідно до розподілу хі-квадрат з $\nu = (r-1)(c-1)$ ступенями свободи.

Критерій Зелтермана має вигляд:

$$D_z^2 = \chi^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}}{E_{ij}} + rc, \qquad (7)$$

де значення χ^2 обчислено, за співвідношенням (1). Статистика цього критерію має χ^2-розподіл з $\nu = (r-1)(c-1)$ ступенями свободи при рівні довірчої ймовірності α. В [2] показано, що обчислювати величину χ^2 слід за співвідношенням:

$$\chi^2 = \frac{N(ad-bc)^2}{\alpha\beta\gamma\delta}. \qquad (8)$$

Для таблиці спряженості 2×2 кількість ступенів свободи дорівнює $\nu = 1$. Більш точний результат обчислення величини χ^2 можна отримати, якщо використовувати його незсунену оцінку, визначену за співвідношенням:

$$\chi^2(незс) = \frac{N\left[\left|ad-bc\right| - \frac{1}{2}N\right]^2}{\alpha\beta\gamma\delta}. \qquad (9)$$

Ентропію дискретної випадкової величини, закон розподілу якої визначено в табл. 1, визначають за добре відомим співвідношенням:

$$H = -\sum_{i=1}^{s}\frac{n_i}{n}\ln\left(\frac{n_i}{N}\right) = -\sum_{i=1}^{s} p_i \ln p_i. \qquad (10)$$

Для визначення незсуненої оцінки $\hat{H}$ ентропії H випадкової величини, гістограма якої отримана за наслідками спостережень, в [16] отримано співвідношення:

$$\hat{H} = -\sum_{i=1}^{s} p_i \ln p_i - \frac{s-1}{N} + \frac{1-\sum_{i=1}^{s} p_i^{-1}}{12N^2} + \frac{\sum_{i=1}^{s}\left(p_i^{-1} - p_i^{-2}\right)}{12N^3}. \qquad (11)$$

Оцінка (11) відрізняється від співвідношення (10) тим, що містить поправку на зсунення. Дисперсію ентропії визначають за співвідношенням:

$$D_H^2 = \frac{1}{N}\left(\sum_{i=1}^{s} p_i \ln^2 p_i - H^2\right); \qquad (12)$$

її незсунену оцінку $\hat{D}_H^2$ визначають як:

$$\hat{D}_H^2 = \frac{1}{N}\left(\sum_{i=1}^{s} p_i \ln^2 p_i - H^2\right) + \frac{s-1}{2N^2}. \qquad (13)$$

Середньоквадратичне відхилення величини ентропії d_H і його оцінку $\hat{d}_H$ визначають за співвідношеннями:

$$d_H = \sqrt{D_H^2}\ ,\ \hat{d}_H = \sqrt{\hat{D}_H^2}\ . \qquad (14)$$

Нижню H_d і верхню H_u межі довірчого інтервалу оцінки ентропії $\hat{H}$ в роботі [16] запропоновано визначати за співвідношеннями:

$$H_d = \hat{H} - \Omega\left(\frac{1+\beta}{2}\right)\cdot\frac{\hat{d}_H}{\sqrt{s}} = \hat{H} - \gamma\cdot\frac{\hat{h}_H}{\sqrt{s}}; \qquad (15)$$

$$H_u = \hat{H} + \Omega\left(\frac{1+\beta}{2}\right)\cdot\frac{\hat{d}_H}{\sqrt{s}} = \hat{H} + \gamma\cdot\frac{\hat{h}_H}{\sqrt{s}}. \qquad (16)$$

У співвідношеннях (15) та (16) прийнято, що $\Omega(\bullet)$ – обернена функція стандартного нормального розподілу, в подальших обчисленнях за умовчанням прийнято, що β=0,975, γ=1,96. В [17] запропоновано обчислювати відносні оцінки ентропії: величину відносної ентропії

$$h = \frac{H}{H\max} = \frac{H}{\ln s} \qquad (17)$$

та інформаційний індекс різноманітності

$$I_{div} = \frac{1}{N}\cdot\ln\left(N!\Big/\prod_{i=1}^{s} n_i!\right). \qquad (18)$$

У [18] для порівняння двох ентропій запропоновано варіант t-критерію Стьюдента у вигляді:

$$t_{факт} = \frac{\left|H_1 - H_2\right|}{\sqrt{\left(D_{H1}^2\right)^2\Big/N_1 + \left(D_{H2}^2\right)^2\Big/N_2}}, \qquad (19)$$

Кількість ступенів свободи визначають як:

$$df = \left[\frac{\left(D_{H1}^2 + D_{H2}^2\right)^2}{\left(D_{H1}^2\right)^2\Big/N_1 + \left(D_{H2}^2\right)^2\Big/N_2}\right], \qquad (20)$$

де $[A]$ – ціла частина числа A.

Методика ентропійного аналізу даних, які представлені в табл. 2, викладена в [8, 13, 19]. Відповідно до рекомендацій цих робіт безумовну ентропію даних визначали за співвідношенням:

$$H(X,Y) = -\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}}{N}\ln\left(\frac{n_{ij}}{N}\right). \quad (21)$$

Ентропійную міру дисперсії визначали за співвідношенням:

$$\varepsilon = \left(N\ln N - \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\ln n_{ij}\right) \Big/ \left(N\ln(rc)\right). \quad (22)$$

Умовні ентропії визначали за співвідношеннями:

$$H_y = -\sum_{j=1}^{c}\frac{n_{\bullet j}}{N}\ln\frac{n_{\bullet j}}{N}; \quad (23)$$

$$H_x = -\sum_{i=1}^{r}\frac{n_{i\bullet}}{N}\ln\left(\frac{n_{i\bullet}}{N}\right); \quad (24)$$

$$H(y/x_i) = -\sum_{j=1}^{c}\frac{n_{ij}}{n_{i\bullet}}\ln\left(\frac{n_{ij}}{n_{i\bullet}}\right),\ i=\overline{1,r}; \quad (25)$$

$$H(x/y_j) = -\sum_{i=1}^{r}\frac{n_{ij}}{n_{\bullet j}}\ln\left(\frac{n_{ij}}{n_{\bullet j}}\right),\ j=\overline{1,c}. \quad (26)$$

Методи ентропійного аналізу дозволяють визначати міру статистичного зв'язку між парою змінних (*X*, *Y*) незалежно від шкал, в яких вони визначені, та не залежать від форми зв'язку між ними і закону їх розподілу. В [20] ця характеристика названа інформаційним коефіцієнтом кореляції. Чисельне значення інформаційного коефіцієнта кореляції *R* (*X*, *Y*) визначають із співвідношень:

$$I = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}}{N}\ln\frac{N\cdot n_{ij}}{n_{i\bullet}\times n_{\bullet j}}; \quad (27)$$

$$R(X,Y) = \sqrt{1-\exp(-2I(X,Y))}. \quad (28)$$

В [21] запропоновано при обчисленні співвідношення (27) робити поправку на зсунення. В цьому випадку слід використовувати співвідношення:

$$\hat{I} = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}}{N}\ln\frac{N\cdot n_{ij}}{n_{i\bullet}\cdot n_{\bullet j}} - \frac{(r-1)(c-1)}{2N}; \quad (29)$$

$$\rho(X,Y) = \sqrt{1-\exp(-2\hat{I}(X,Y))}. \quad (30)$$

Порівняння інформаційного коефіцієнта кореляції з коефіцієнтами кореляції Пірсона і Спірмена – одна із задач цієї роботи.

В [13] для порівняння гістограм розподілу об'єктів, властивості яких визначено в однакових шкалах та які мають однакову кількість інтервалів, описано використання відстаней Хелінгера і Кулльбака-Леблера. Властивості і особливості визначення відстані Хелінгера для гістограм, які приведені в табл. 4, описано в [22, 31].

Таблиця 4 – **Загальний вид порівнюваних гістограм**

Гістограма закону розподілу випадкової величини P					
n_1^P	n_2^P	·	n_i^P	…	n_s^P
p_1	p_2	·	p_i	…	p_s
Гістограма закону розподілу випадкової величини Q					
n_1^Q	n_2^Q	…	n_i^Q	…	n_s^Q
q_1	q_2	…	q_i	…	q_s

Загальна кількість спостережень в кожній з гістограм визначена співвідношеннями:

$$\sum_{i=1}^{s} n_i^P = N_P;\quad \sum_{i=1}^{s} n_i^Q = N_Q. \quad (31)$$

Частка спостережень у кожному з інтервалів гістограм визначають за співвідношеннями:

$$p_i = n_i^P / N_P;\quad q_i = n_i^Q / N_Q. \quad (32)$$

Відстань Хелінгера між гістограмами *P* і *Q* визначають за співвідношенням, яке наведено в [31]:

$$D_H(P,Q) = \\ = \frac{1}{\sqrt{2}}\left\|\sqrt{P}-\sqrt{Q}\right\| = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{s}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2}. \quad (33)$$

Для гістограм, які наведені в табл. 4, відстань Кулльбака-Леблера визначають за співвідношеннями:

$$D_{KL}(P,Q) = \sum_{i=1}^{s} p_i\ln\left(\frac{p_i}{q_i}\right); \\ D_{KL}(Q,P) = \sum_{i=1}^{s} q_i\ln\left(\frac{q_i}{p_i}\right). \quad (34)$$

З (34) витікає, що відстань Кулльбака-Леблера має важливу особливість – воно не симетричне, тобто $D_{KL}(P,Q) \neq D_{KL}(Q,P)$. Приклад застосування цієї відстані для кількісної міри оцінки відмінності між гістограмами різних розподілів приведено в [24]. В [25] запропонована методика перевірки статистичних гіпотез, яка заснована на використанні відстані Кулльбака-Леблера. Для викладу її основних положень в співвідношеннях (36)…(41) і далі будуть використані терміни і позначення, що прийняті в цій роботі. Задачею однієї вибірки для поліноміальної популяції в [25] названа задача в наступній постановці. Припустимо, що результати спостережень представлено у вигляді гістограми, наведеної в табл. 5. У цій таблиці прийнято, що в кожному *i*-ому інтервалі гістограми, $i=\overline{1,s}$, міститься x_i спостережень. Прийнята для аналізу отриманої гістограми статистична модель дозволяє визначити величину p_i – розрахункову частку кількості спостережень в кожному інтервалі за умови, що:

$$\sum_{i=1}^{s} x_i = N;\ \sum_{i=1}^{s} p_i = 1. \quad (35)$$

Таблиця 5 – **Умови задачі однієї вибірки**

N	x_1	x_2	…	x_i	…	x_s
P	p_1	p_2	…	p_i	…	p_s

Необхідно перевірити гіпотезу H_0 , яка полягає в тому, що розподіл спостережень випадкової величини N не протирічить теоретичному розподілу випадкової величини P. В цьому випадку гіпотеза H_1 – альтернативна гіпотеза. В такій постановці задача співпадає з перевіркою гіпотези щодо закону розподілу випадкової величини по критерію χ^2. Для розв'язання поставленої задачі в [25] пропонується обчислити дві величини:

$$\hat{I}(^*:2;O_N)=\sum_{i=1}^{s}x_i\ln\frac{x_i}{Np_i}; \qquad (36)$$

$$\hat{J}\left(^*:2;O_N\right)=\left[\sum_{i=1}^{s}\left(\frac{x_i}{N}-p_i\right)\ln\frac{x_i}{Np_i}\right], \qquad (37)$$

які мають розподіл χ^2 з (s - 1) ступенем свободи.

Задачею двох вибірок для полиноміальної популяції названа задача в наступній постановці. Для гістограм, які наведено в табл. 4, необхідно перевірити гіпотезу H_0 про те, що ці розподіли належать одній популяції. Для розв'язання поставленої задачі в [25] запропоновано обчислити величини:

$$2\hat{I}(H_1:H_0)\approx\frac{1}{N_P\ N_Q}\sum_{i=1}^{s}\frac{\left(N_Q n_i^P-N_P n_i^Q\right)^2}{n_i^P+n_i^Q}; \qquad (38)$$

$$\hat{J}\left(H_1:H_0\right)=\frac{1}{2\left(N_p+N_q\right)^2}+$$

$$+\frac{1}{2N_PN_Q}\sum_{i=1}^{s}\frac{\left(N_Q n_i^P-N_P n_i^Q\right)^2}{n_i^P+n_i^Q}+ \qquad (39)$$

$$+\sum_{i=1}^{s}\frac{\left(N_Q n_i^P-N_P n_i^Q\right)^2\left(n_i^P+n_i^Q\right)}{n_i^P\,n_i^Q},$$

які мають розподіл χ^2 з (s - 1) ступенем свободи.

Для таблиць спряженості типів 2×2 (табл. 2) і $r\times c$ (табл. 3) при перевірці гіпотези H_0, яка полягає в перевірці відсутності зв'язку між ефектами рядків і стовпців (альтернативна гіпотеза H_1 – присутній зв'язок між ефектами рядків і стовпців), слід обчислити співвідношення:

$$2\hat{I}(H_1:H_0)=\sum_{i=1}^{r}\sum_{i=1}^{c}\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{N}\right)^2\Bigg/\frac{n_{i\bullet}n_{\bullet j}}{N}; \qquad (40)$$

$$\hat{J}(H_1:H_0)=\frac{1}{2}\left[\begin{array}{l}\displaystyle\sum_{i=1}^{r}\sum_{i=1}^{c}\frac{\left(n_{ij}-n_{i\bullet}n_{\bullet j}/N\right)^2}{n_{i\bullet}n_{\bullet j}/N}+\\ \displaystyle+\sum_{i=1}^{r}\sum_{i=1}^{c}\frac{\left(n_{ij}-n_{i\bullet}n_{\bullet j}/N\right)^2}{n_{ij}}\end{array}\right]. \qquad (41)$$

В [25] показано, що чисельні значення співвідношень (40) і (41) мають розподіл χ^2 з $(r - 1)(c - 1)$ ступенями свободи. Гіпотези, які сформульовані за співвідношеннями (36, 37), (38, 39), (40, 41), приймаються тільки при сумісному виконанні кожної з гіпотез, що входить у відповідну пару.

Приклади застосування відстані Кулльбака-Леблера для аналізу фінансової інформації і рішення задач проектування критичних систем описані в [27, 28, 29, 31]. При підготовці даного повідомлення автори не змогли знайти відомості про програмні продукти, що реалізують ентропійні методи аналізу таблиць спряженості.

Мета роботи. Розробка EXCEL-орієнтованого калькулятора для обчислення результатів ентропійного аналізу даних, які розподілені по категоріях.

Предмет дослідження – гістограми довільних законів розподілу та таблиці спряженості 2×2.

Методи дослідження: Ентропійний та інформаційний аналіз гістограм довільних законів розподілу та таблиць спряженості 2×2.

Отримані результати

На думку авторів даного повідомлення, систему EXCEL можна розглядати як одну з найбільш зручних платформ для обробки даних про функцію щільності ймовірностей, які представлено в табличній формі, тобто у вигляді гістограми. В даній роботі процес обробки, даних побудовано на використанні "розумної таблиці" – інструменту, який дозволяє спростити формули з посиланнями на її елементи. Головною перевагою "розумної таблиці" є те, що при зміні кількості рядків у ній формули автоматично налаштовуються на ці зміни. Крім початкових даних таблиця містить стовпці з формулами для обчислення проміжних результатів.

На головному аркуші робочої книги Excel розміщується кнопка виклику меню програми, яке містить перелік задач (рис. 1). Як вікно меню, так і вікна з результатми задач створено за допомогою форм – стандартного інструменту MS Excel, що значно спрощує режим діалогу. Для виклику на екран результатів будь-якої задачі треба її виділити і натиснути кнопку "Пуск", розміщену у вікні меню праворуч. При цьому вікно меню з екрана вилучається, а на його місці виводиться вікно з результатами відповідної задачі. Після цього знову можна викликати вікно меню і отримати результати іншої задачі. Вікна з результатми різних задач можна розмістити на екрані поруч одне з оним, або закривати непотрібні.

Перший пункт меню відрізняється від інших задач, бо здійснює перехід на аркуш з таблицею, у яку треба ввести початкові дані. Введення у таблицю нових даних складається у заповненні стовпця "Кількість елементів n_i". Якщо нових даних більше, ніж у існуючої таблиці, розрахункові формули автоматично додаються у нові рядки. Якщо в таблиці виявляються зайві рядки, то їх треба вилучити командою "вилучити комірки". Варіант застосування інструмента "Розумна таблиця" показано на рис. 2.

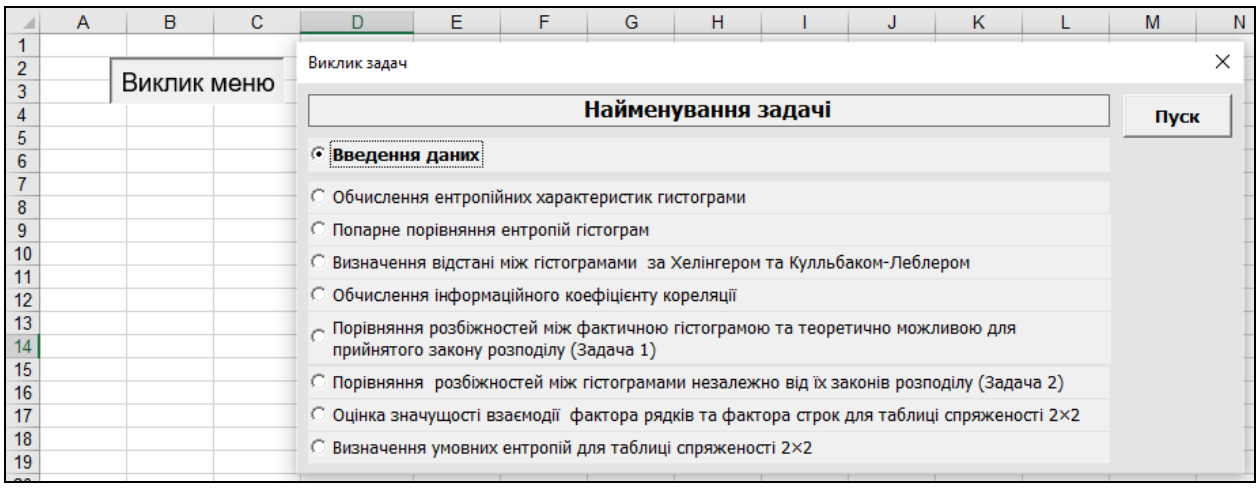

Рис. 1. Скрін-копія аркуша таблиці з меню програми
(**Fig.1.** Screen – a copy of the screen window for the table sheet from the program menu)

№ Класу	Кількість елементів n_i	$P_i=n_i/\Sigma n_i$	$Ln(P_i)$	$P_i ln(P_i)$	$ln(P_i)^2$	$P_i\, ln(P_i)^2$	корінь(Pi)
1	10	0,058823529	-2,833213344	-0,166659608	8,027097853	0,472182227	0,242535625
2	20	0,117647059	-2,140066163	-0,25177249	4,579883184	0,538809786	0,34299717
3	50	0,294117647	-1,223775432	-0,35993395	1,497626307	0,440478326	0,542326145
4	35	0,205882353	-1,580450376	-0,325386842	2,49782339	0,514257757	0,453742606
5	28	0,164705882	-1,803593927	-0,297062529	3,252951053	0,535780173	0,405839725
6	15	0,088235294	-2,427748236	-0,21421308	5,893961497	0,520055426	0,297044263
7	12	0,070588235	-2,650891787	-0,187121773	7,027227268	0,496039572	0,265684466

Рис. 2. Скрін-копія вікна екрану варіанта застосування функції "Розумна таблиця"
(**Fig. 2.** Screen – a copy of the screen window of the option to use the "Dynamic Table" function)

Для виконання ентропійного аналізу одновимірних гістограм їх представляють у вигляді, який показано в табл. 6.

Таблиця 6 – **Приклад подання гістограм**

Гісто-грами	Комірки гістограм							Усього спо-стережень
	1	2	3	4	5	6	7	
Г1	10	20	50	35	28	15	12	170
Г2	24	24	24	25	25	24	24	170

Калькулятор обробляє одну, декілька або всі введені гістограми згідно з вказівкою користувача. Для прикладу застосування калькулятора обрано гістограму Г1, яку наведено в [4], та гістограму Г2, що відповідає рівномірному закону розподілу. Результати обчислення ентропійних характеристик гістограми Г1 показано на рис. 3. Для їх визначення було використано співвідношення (10)...(18).

Для попарного порівняння ентропій гістограм Г1 (*H1)* і Г2 (*H2)* використано критерій Хатчинсона, який визначений співвідношеннями (19), (20). Як нульова прийнята гіпотеза *H0*: *H1=H2*, як альтернативна прийнята гіпотеза *H1*: $H1 \neq H2$. У цьому випадку правила для прийняття гіпотез мають вигляд:

$$T_3 = \begin{cases} H_0, & \text{якщо} \quad Pv(t_\phi) > Pv_\alpha; \\ H_1, & \text{якщо} \quad Pv(t_\phi) < Pv_\alpha. \end{cases} \qquad (42)$$

У співвідношенні (42) величину $Pv_\phi(t)$ визначають, використовуючи умову:

$$Pv_\phi(t) := \text{СТЬЮДРАСП}\,(t_\phi; df; 2). \qquad (43)$$

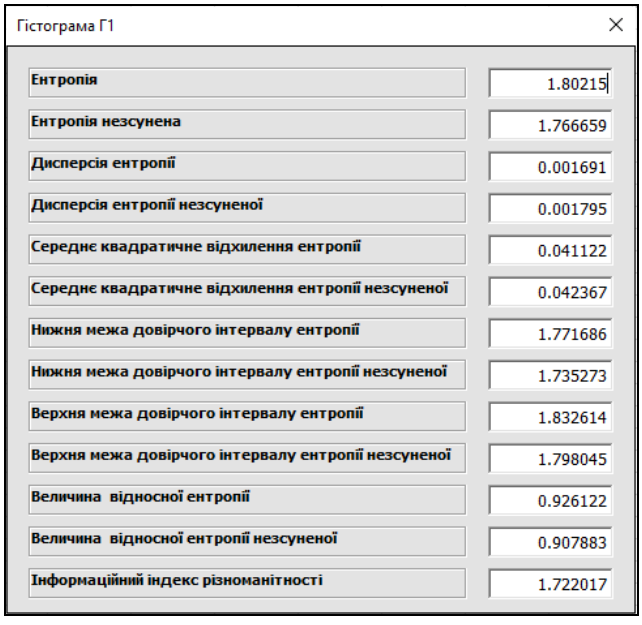

Рис. 3. Скрін-копія вікна екрану обчислення ентропійних характеристик гистограми Г1
(**Fig. 3.** Screen – a copy of the screen window for calculating the entropy characteristics of the histogram Г1)

Величину t_ϕ визначали, використовуючи співвідношення (19), величину *df* – (20). За замовченням величина α=0,05. Результати порівняння гістограм Г1 і Г2 показані на рис. 4. Ентропія H1 відповідає гістограмі Г1, ентропія H2 відповідає гістограмі Г2. Всі обчислення виконують для зсуненої та незсуної оцінок гістограм.

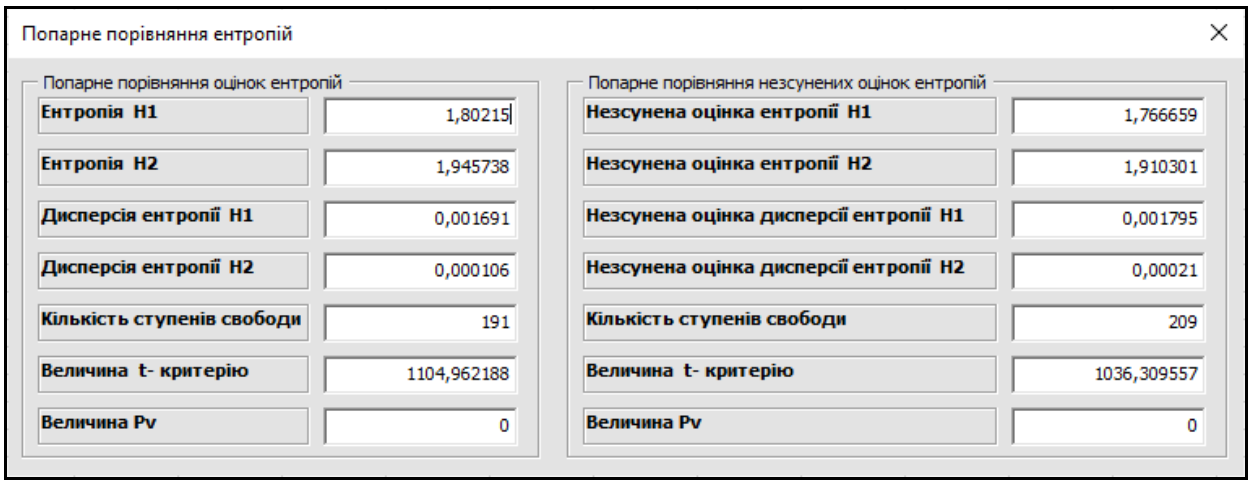

Рис. 4. Скрін-копія вікна екрану попарного порівняння ентропій гістограм Г1 (*H1*) и Г2 (*H2*)
(**Fig. 4.** Screen – a copy of the window of the screen for pairwise comparison of the entropies of the histograms Г1 (*H1*) and Г2 (*H2*))

Як і слід було чекати, за наслідками порівняння слід прийняти альтернативну гіпотезу H1. Це не протирічить фізичному змісту отриманого результату тому,що гістограми Г1 і Г2 відповідають абсолютно різним законам розподілу. В рамках даної роботи прийняття статистичних гіпотез, пов'язане з використанням величини χ^2, виконують згідно із співвідношенням (4), обгрунтування використання величини *Pv* та методи її визначення наведено в [37]. Для аналізу таблиць спряженості 2×2 цю величину визначають за співвідношенням:

$$Pv\left(\chi^2\right) := \text{ХИРАСП}(\chi^2_{\phi});1). \qquad (44)$$

Для обчислення відстані між парами гістограм використовували відстань Хелінгера, яку визначали згідно із співвідношеннями (31)…(33) і відстань Кулльбака-Леблера, яку визначали згідно із співвідношеннями (31), (32), (34). Результати обчислення показано на рис. 5.

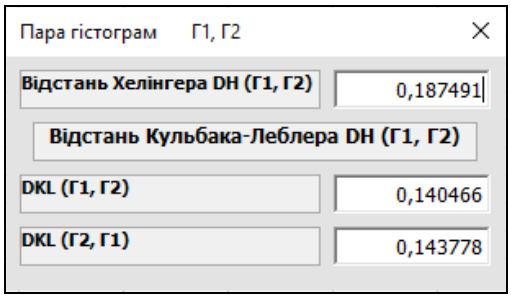

Рис. 5. Скрін-копія вікна екрану визначення відстані між гістограмами Г1 и Г2
(**Fig. 5.** Screen – a copy of the screen window for determining the distance between histograms Г1 and Г2)

До методів ентропійного аналізу відносять також і обчислення інформаційного коефіцієнта кореляції. Способи його обчислення описано в [20, 21]. Для порівняння інформаційного коефіцієнта кореляції, обчисленого за співвідношеннями (27) та (28) з коефіцієнтом кореляції Пірсона було проведено чисельний експеримент. В роботах [32…34] наведено кореляційні таблиці і значення коефіцієнтів кореляції Пірсона або інформаційного коефіцієнта кореляції. Для кожної з цих таблиць (табл. 7 … 9) було обчислено недостатні коефіцієнти. У комірках цих таблиць поставлено відповідну кількість спостережень.

Для якісного оцінювання величини коефіцієнту кореляції використаємо так звану шкалу Чеддока (1879-1940), табл. 10. Незважаючи на те, що цю шкалу застосовують у багатьох роботах, наприклад [30], посилання на першоджерело як правило відсутні. Автори цього повідомлення із задоволенням заповнюють цю прогалину і відсилають читачів на роботу [35].

Результати обчислень показано у табл. 11.

Таблиця 7 – **Кореляційна таблиця, побудована за даними роботи [32]**

Умовні індекси змінної *X*	**Умовні індекси змінної *Y***							
	A	B	C	D	E	F	G	H
1							1	
2							1	
3						1		
4						1	1	1
5					2	1		
6				1	1	4		
7			1	4	2			
8	1	3	9		3			
9		2	7	1	1			
10		1	5		1			
11			2		1	1		

Таблиця 8 – **Кореляційна таблиця, побудована за даними роботи [33]**

Умовні індекси змінної *X*	**Умовні індекси змінної *Y***							
	A	B	C	D	E	F	G	H
1		1						
2	1	6						
3		5	9					
4		4	9	5	3			
5			7	9	3			
6			2	9	12	4	2	
7				7	8	6	4	
8				1	2	5	3	3
9						1	2	2

Таблиця 9 – **Кореляційна таблиця, побудована за даними роботи [34]**

Умовні індекси змінної *X*	Умовні індекси змінної *Y*							
	A	B	C	D	E	F	G	H
1	1	5	7	4				
2			1	5	2			
3			1	1	6	3	2	
4					2	4	2	
5					1	1		

Таблиця 10 – **Якісна оцінка величини коефіцієнта кореляції (шкала Чеддока)**

Значення коефіцієнта кореляції	[0,1...0,3)	[0,3...0,5)	[0,5…0,7)
Зв'язок	незначний	помірний	істотний
Рівень	1	2	3

Значення коефіцієнта кореляції	[0,7...0,9)	[0,9...0,99]	1,0
Зв'язок	високий	дуже високий	Функціональний
Рівень	4	5	6

Таблиця 11 – **Чисельні значення коефіцієнту кореляції Пірсона та інформаційного коефіцієнту кореляції**

Показники вимірів Статистичного зв'язку•)	Літературні посилання		
	[32]	[33]	[34]
$\hat{I}$	0,884	0,862	0,876
ρ	0,547**(3)**	0,782**(4)**	0,730**(4)**
r	0,669 **(4)**	0,832**(4)**	0,859**(4)**

•)Жирним шрифтом позначена якісна оцінка величини коефіцієнта кореляції за шкалою Чеддока (табл. 10).

З цієї таблиці випливає, що інформаційний коефіцієнт кореляції оцінює на якісному рівні зв'язок між випадковими величинами так само, як і традиційний коефіцієнт кореляції Пірсона. Для оцінки застосування інформаційного коефіцієнта кореляції при аналізі таблиць спряженості був проведений чисельний експеримент, в процесі якого для отриманих випадковим чином таблиць обчислювали критерії, визначені співвідношеннями (5), (7), (8), (9), (27), (28). Детальний виклад способу отримання цих таблиць у цій роботі не розглянуто. Для табл.11 з урахуванням позначень, які наведено в табл. 3, кількість інформації визначали, використовуючи співвідношення:

$$I = \frac{a}{N}\ln\frac{aN}{\alpha\lambda} + \frac{b}{N}\ln\frac{bN}{\alpha\delta} + \frac{c}{N}\ln\frac{cN}{\gamma\beta} + \frac{d}{N}\ln\frac{dN}{\beta\delta}. \quad (45)$$

Вихідні дані для експерименту, враховуючи позначення, прийняті в табл. 3, наведені у табл. 12.

Таблиця 12 – **Елементи таблиць спряженості 2×2, отримані випадковим методом**

№ таблиць	Елементи таблиць				№ таблиць	Елементи таблиць			
	a	b	c	d		a	b	c	d
1	17	23	7	32	11	25	14	17	24
2	27	31	11	10	12	24	7	29	20
3	19	24	22	16	13	22	10	21	27
4	15	32	11	22	14	9	34	6	38
5	12	14	25	28	15	40	11	13	16
6	22	10	33	16	16	40	6	8	26
7	5	17	19	39	17	15	26	25	14
8	8	15	17	40	18	30	14	6	31
9	24	22	10	25	19	30	19	5	26
10	12	17	18	33	20	9	18	28	25

Чисельні значення критеріїв, які використовували при аналізі таблиць спряженості, наведено в табл. 13.

Оскільки значення критерію Цоя-Самена менше величини 0,25, то виконання подальшого аналізу критеріїв (I) … (V) вважатимуться коректним. Для визначення зв'язку між цими критеріями побудована матриця їх взаємної кореляції, яка наведена в табл. 14.

Таблиця 13 – **Чисельні значення статистичних критеріїв, які було використано при аналізі випадкових таблиць спряженості 2×2**

№ таблиці	Критерій хі-квадрат (I) ••)		Критерій хі-квадрат незсунений(II) ••)		Критерій відношення gравдоподібності (III) ••)		Критерій Зелтермана (IV) ••)		Информаційний коефіцієнт кореляції (V) ••)	Критерій Цоя-Самена
	Чисельне значення	Величина Pv	Чисельне значення	Величина Pv	Чисельне значення	Величина Pv	Чисельне значення	Величина Pv	Чисельне значення	Чисельне значення
1	5,627	0,012	4,531	0,035	5,763	0,016	5,633	0,017	0,265	0,09
2	0,209	0,646	0,04	0,839	0,209	0,646	0,205	0,65	0,051	0,1
3	1,516	0,218	1,02	0,313	1,521	0,217	1,517	0,218	0,136	0,07
4	0,017	0,893	0,012	0,913	0,017	0,893	0,013	0,906	0,015	0,09
5	0,07	0,932	0,02	0,876	0,007	0,932	0,008	0,017	0,010	0,08
6	0,017	0,894	0,01	0,911	0,017	0,894	0,022	0,281	0,015	0,09
7	0,764	0,381	0,36	0,549	0,791	0,373	0,85	0,356	0,099	0,124
8	0,187	0,665	0,03	0,867	0,185	0,666	0,15	0,697	0,048	0,117
9	4,546	0,03	3,63	0,057	4,632	0,031	4,567	0,032	0,236	0,07
10	0,292	0,588	0,09	0,76	0,29	0,589	0,274	0,6	0,060	0,089

Закінчення табл. 13

11	4,107	0,042	3,25	0,071	4,146	0,041	4,109	0,042	0,225	0,073
12	2,823	0,093	2,07	0,15	2,914	0,087	2,883	0,089	0,189	0,092
13	4,827	0,028	3,87	0,049	4,913	0,026	4,842	0,027	0,244	0,078
14	0,81	0,367	0,38	0,537	0,814	0,366	0,806	0,369	0,097	0,132
15	9,336	0,002	7,89	0,005	9,224	0,002	9,202	0,002	0,330	0,095
16	32,77	$<1\cdot10^{-3}$	30,18	$<1\cdot10^{-3}$	34,95	$<1\cdot10^{-3}$	32,69	$<1\cdot10^{-3}$	0,595	0,08
17	6,053	0,014	5	0,025	6,132	0,014	6,053	0,014	0,272	0,073
18	21,98	$<1\cdot10^{-3}$	19,93	$<1\cdot10^{-3}$	23,44	$<1\cdot10^{-3}$	22,001	$<1\cdot10^{-3}$	0,272	0,074
19	15,69	$<1\cdot10^{-3}$	14,01	0,002	16,82	$<1\cdot10^{-3}$	15,741	$<1\cdot10^{-3}$	0,435	0,08
20	2,735	0,098	2,01	0,1562	2,777	0,095	2,75	0,096	0,185	0,086

••) Умовні індекси критеріїв

Таблиця 14 – **Коефіцієнти взаємної кореляції між критеріями, які використані при аналізі таблиць спряженості 2×2**

Умовні індекси критеріїв	Умовні індекси критеріїв				
	I	II	III	IV	V
I	1	0,999	0,995	0,993	0,987
II		1	0,995	0,992	0,986
III			1	0,998	0,992
IV				1	0,993
V					1

З даних, наведених у цій таблиці, випливає, що інформаційний критерій кореляції рядків та стовпців таблиці спряженості може бути обґрунтовано використаний при аналізі цих таблиць. Перевага його перед іншими критеріями в тому, що він не тільки визначає наявність (відсутність) статистичного зв'язку між факторами рядків та стовпців, а й дає її кількісну оцінку. Скрін-копія вікна екрана з результатами обчислення інформаційного коефіцієнту кореляції згідно із (27), (28) показана на рис. 6.

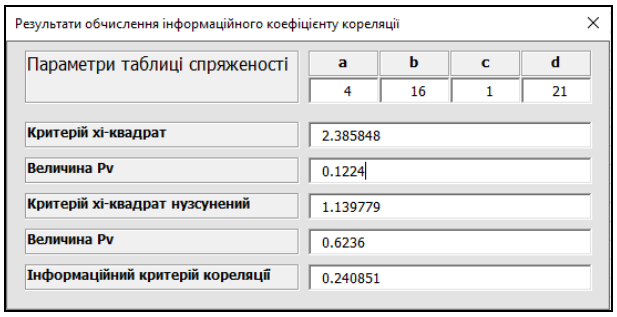

Рис. 6. Скрін-копія вікна екрану з результатами обчислення інформаційного коефіцієнту кореляції
(**Fig. 6.** Screen – a copy of the screen window with the results of the information correlation coefficient calculation)

Приклади розв'язання задач визначення значущості розбіжності між гістограмами згідно із співвідношеннями (36)…(39) наведено на рис. 7, 8.

Приклад оцінки значущості взаємодії фактора рядків та фактора стовпчиків для таблиці спряженості 2×2 згідно із співвідношеннями (40) та (41) прказаний на рис. 9.

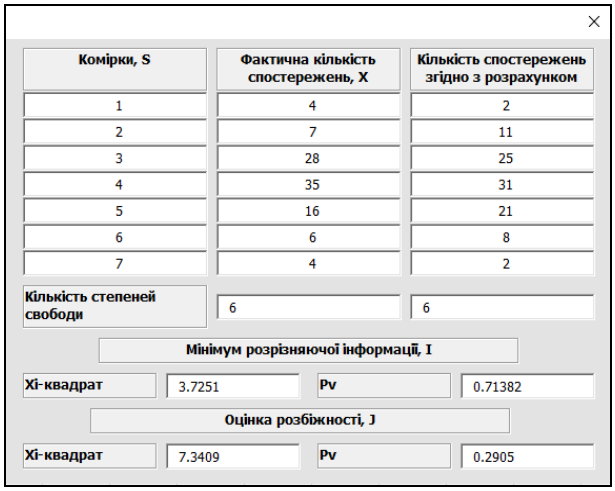

Рис. 7. Скрін-копія вікна екрану з результатами порівняння розбіжностей між фактичною гістограмою та теоретично можливою для прийнятого закону розподілу, Задача 1
(**Fig. 7.** A screenshot of the screen copy of the results of the difference between the actual histogram and the theoretically possible for the adopted law of the difference, Task 1)

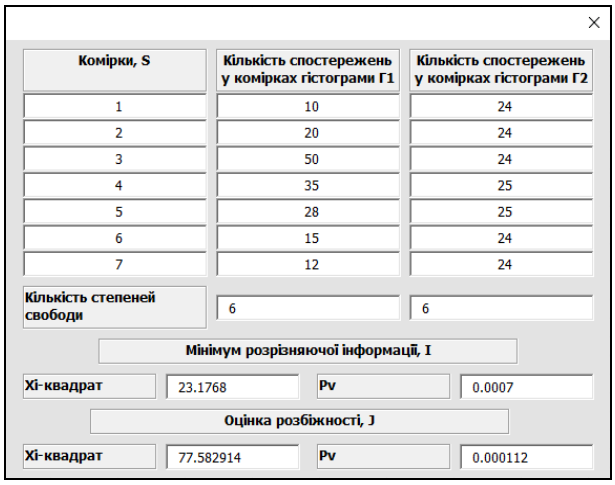

Рис. 8. Скрін-копія вікна екрану з результатами порівняння розбіжностей між гістограмами незалежно від їх законів розподілу, Задача 2
(**Fig. 8.** A screen copy of the screen window with the results of comparing the differences between histograms regardless of their distribution laws, Task 2)

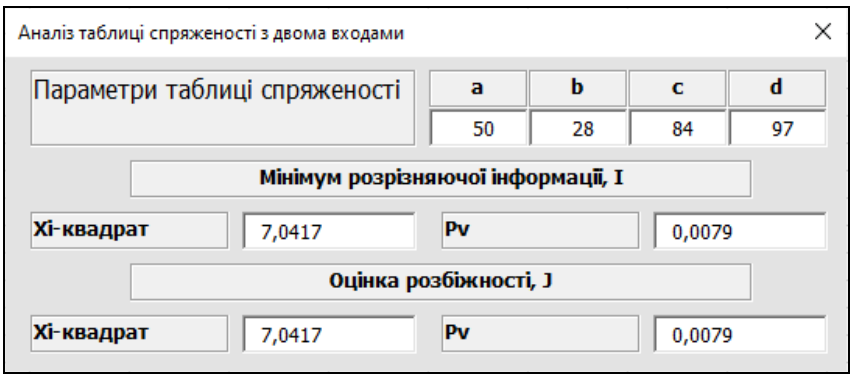

Рис. 9. Скрін-копія вікна екрану з результатами значущості взаємодії фактора рядків та фактора стовпчиків для таблиці спряженості 2×2
(**Fig. 9.** Screen copy of the screen window with the results of the significance of the interaction of the row factor and the term factor for the contingency table 2×2)

Приклад визначення умовних ентропій для таблиці спряженості 2×2 показаний на рис. 10. Прийняті в цьому прикладі позначення та співвідношення, за якими їх розраховували, наведено в табл. 15.

Приклад визначення умовних ентропій для таблиці спряженості 2×2 показаний на рис. 10. Прийняті в цьому прикладі позначення та співвідношення, за якими їх розраховували, наведено в табл. 15.

Таблиця 15 – **Позначення** умовних **ентропій та співвідношення, прийняті для їх розрахунку**

Показники ентропії таблиці спряженості 2×2	**Визначення у вікні калькулятора**	**Розрахункове співвідношення**
Безумовна ентропія, $H(X,Y)$	$H(X,Y)$	(21)
Ентропійна міра дисперсії, ε	ε	(22)
Умовна ентропія, H_y	$\{H(Y)\}$	(23)
Умовна ентропія, H_x	$\{H(X)\}$	(24)
Умовна ентропія H[$Y(y/x_1)$]	$H(y/x_1)$	(25)
Умовна ентропія H[$Y(y/x_2)$]	$H(y/x_2)$	(25)
Умовна ентропія H[$Y(x/y_1)$]	$H(x/y_1)$	(26)
Умовна ентропія H[$Y(x/y_2)$]	$H(x/y_2)$	(26)

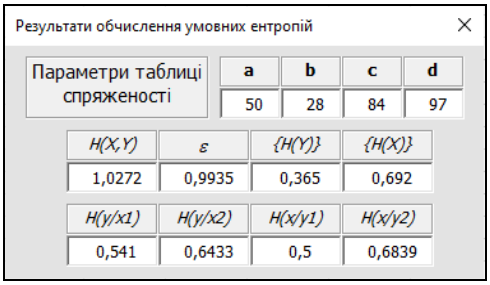

Рис. 10. Скрін-копія вікна екрану з результатами визначення умовних ентропій для таблиці спряженості 2×2
(**Fig. 10** Screen copy of the screen window with the results of determining the conditional entropies for the contingency table)

Висновки

1. Запропоновано використовувати методи ентропійного аналізу для аналізу даних, що розподілені по категоріях та наведено відомості про структуру Excel-орієнтованого калькулятора, призначеного для виконання таких розрахунків.

2. Калькулятор дає можливість обчислювати ентропійні характеристики гістограм, виконувати попарне порівняння ентропій гістограм, визначати відстань між гістограмами, обчислювати інформаційний коефіцієнт кореляції, порівнювати розбіжності між гістограмами.

3. Для таблиць спряженості 2×2 калькулятор дає можливість оцінювати значущість взаємодії фактору рядків та фактору стовпчиків.

4. Калькулятор визначає значення умовних ентропій для таблиць спряженості 2×2.

5. Запропонований калькулятор у деякій мірі заповнює прогалини в існуючих програмних продуктах та може бути використаний для обробки методами ентропійного аналізу даних, що розподілені по категоріях.

6. В роботі показано, що ентропійні методи аналізу гістограм доцільно використовувати у випадках, коли гістограми визначають довільні закони розподілу.

REFERENCES

1. Motalo, V. (2015), “Analysis of measurement scales”, *Measuring technique and metrology*, 2015, No. 76, pp. 21-35, available at: http://nbuv.gov.ua/UJRN/metrolog_2015_76_4.
2. Kendall, Maurice G. and Stuart, Alan (1961). *The Advanced Theory of Statistics. Vol. 2, Inference and Relationship*. Charles Griffin, London, 676 p.
3. Dubnitsky, V. Yu., Kobylin, A. M. and Kobylin, O. A. (2018), “Estimation of the lower bound of the reliability of a physically realizable system during its operation under arbitrary distribution laws of the generalized load and strength”, *Information processing systems*, 2018, No. 1(152), pp. 53-60, doi: https://doi.org/10.30748/soi.2018.152.08.

4. Zhluktenko, V. I., Nakonechnyi, S. I. and Savina, S. S. (2001), "*Probability Theory and Mathematical Statistics. Part II. Mathematical statistics*, KNEU, Kyiv, 336 p., available at: https://www.studmed.ru/zhluktenko-v-nakonechniy-s-savna-ss-teorya-ymovrnostey-matematichna-statistika-u-2-h-ch-ch-matematichna-statistika_3976c660ed4.html.
5. Kulldorff, Gunnar (1961), *Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples*. Almqvist & Wiksell / John Wiley & Sons, Stokholm, 176 p., available at: https://www.amazon.com/Contributions-Estimation-Grouped-Partially-Samples/dp/B0010VDR26.
6. Jun I. V. (1993), "On the number of gradations of histograms of errors in astronomical observations", *Kinematics and physics of celestial bodies*, No. 1, vol. 9, pp. 88-92, available at: https://www.mao.kiev.ua/biblio/jscans/kfnt/1993-09/kfnt-1993-09-1-11.pdf.
7. Jun, I. V. "Mathematical processing of astronomical and space information with non-Gaussian observation errors: Abstract of the thesis. for the competition uch. doctorate degrees. Phys.-Math. sciences: spec. 01.03.01 "Astrometry and Celestial Mechanics", Kyiv, GAO NAS of Ukraine, 1992, 46 p., available at: https://issuu.com/blindguardian/docs/asd.
8. Paniotto, V. I., Maksymenko, V. S. and Kharchenko, N.M. (2004), *Statistical analysis of sociological data*, KM Academy, Kyiv, 2004, 270 p.
9. Tsvetkov, O. V. (2015), *Entropy analysis of data in physics, biology, and technology*], LETI, SPb, 202 p., available at: https://www.researchgate.net/profile/Oleg-Tsvetkov/publication/331686300_entropijnyj_analiz_dannyh_v_fizike_biologii_i_tehnike/links/5c87f3afa6fdcc38174f8a14/entropijnyj-analiz-dannyh-v-fizike-biologii-i-tehnike.pdf.
10. Dubnickij, V. Ju., Filatova, L. D. and Khodyrev, A. I. (2017), "The stability of the estimate of the entropy of the histogram of a continuous random variable with respect to the change in the number of its intervals", *Control, Navigation and Communication Systems*, No 5 (45), pp. 42-46, available at: http://nbuv.gov.ua/UJRN/suntz_2017_5_12.
11. Agresti, A. (2002), *Categorical data analysis*, John Wiley & Sons Inc., New York, 742 p., available at: https://onlinelibrary.wiley.com/doi/book/10.1002/0471249688.
12. Joseph L., Fleiss, Bruce, Levin and Myunghee Cho, Paik (2003), *Statistical Methods for Rates and Proportions*, John Wiley & Sons, Inc. New York, 768 p., available at: https://onlinelibrary.wiley.com/doi/book/10.1002/0471445428.
13. Nobuoki, Eshima (2020), *Statistical Data Analysis and Entropy*, Springer Nature Singapore Pte Ltd, Singapore, 498 p., available at: https://link.springer.com/book/10.1007/978-981-15-2552-0.
14. Graham J. G., Upton (1978), *The Analysis of Cross-tabulated Data*, J. Wiley, New York, 160 p., available at: https://www.amazon.com/Analysis-Cross-tabulated-Data-Graham-Upton/dp/0471996599.
15. Duncan, Crammer (2003), *Advanced Quantative Data Analysis*, Open University Press, Philadelphia, 272 p., available at: https://www.amazon.com/Advanced-Quantitative-Analysis-Understanding-Research/dp/0335200591.
16. Anne E., Magurran (1983), *Ecological Diversity and its Measurement*, London, Sydney, CROOM HELM Royal Society University Research Fellow University, 184 p., available at: https://link.springer.com/book/10.1007/978-94-015-7358-0.
17. Margalef, R. (1958), "Information theory in ecology", *Gen. Syst.*, No 3, pp. 36-71, available at: https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1134401.
18. Hutcheson, K. (1970), "A Test for Comparing Diversities Based on the Shannon Formula", *Journal of Theoretical Biology*, vol. 29, pp. 151- 154, doi: http://dx.doi.org/10.1016/0022-5193(70)90124-4.
19. Michalowicz, J.V., Nichols, J.M. and Bucholtz, F. (2014), *Handbook of differential entropy*, Taylor & Francis Group, LLC, London, 241 p., available at: https://www.routledge.com/Handbook-of-Differential-Entropy/Michalowicz-Nichols-Bucholtz/p/book/9781138374799.
20. Linfoot, E. and Linfoot, E. H. (1957), "An Informational Measure of Correlation", *Information and Control*, vol. 1, No. 1, pp. 85-89, available at: https://www.sciencedirect.com/science/article/pii/S001999585790116X.
21. Vistelius, A. B. (1960), "The skew frequency distributions and the fundamental law of Geochemical processes", *The Journal of Geology*, vol. 68, No. 1, pp. 1-22, available at: https://www.jstor.org/stable/30058252.
22. Marian, P. and Marian, T. A. (2015), "Hellinger distance as a measure of Gaussian discord", *The Journal of Physics A: Math. Theor*., 48:11 (2015), 115301, 21 p., arXiv: 1408.4477, doi: http://dx.doi.org/10.1088/1751-8113/48/11/115301.
23. Kullback, S. and Leibler, R.A. (1951), "On information and sufficiency", *The Annals of Mathematical Statistics*, Vol. 22. No. 1. pp. 79-86, doi: http://dx.doi.org/10.1214/aoms/1177729694.
24. Dubnytskyi, V. Yu., Skorykova, I. G. and Khodyrev, O. I. (2017), "Optimal approximation of the distribution density function according to the minimum information loss criterion", *Information processing systems*, No. 4, pp. 45-51, available at: https://www.hups.mil.gov.ua/periodic-app/article/17653.
25. Solomon, Kullback (1978), *Information Theory and Statistics*, Peter Smith, Gloucester, Mass, 399 p., available at: https://books.google.com.ua/books/about/Information_Theory_and_Statistics.html?id=XeRQAAAAMAAJ&redir_esc=y.
26. Hadetska, S. V., Dubnytskyi, V. Yu., Kushneruk, Yu. I. and Hodyrev, O. I. (2020), "A specialized software calculator for evaluating the clinical informativeness of laboratory tests", *Advaned Inforvation Systems*, No. 2, Vol. 4, pp. 80-84, doi: https://doi.org/10.20998/2522-9052.2020.2.12.
27. Soloshenko, O. M. (2014), "Study of the Kullback-Leibler distance in modeling problems in credit scoring", *Development of information-resource support for education and science in the mining and metallurgical industry and in transport, September 27-28, 2014*, Dnepropetrovsk, pp. 328-333, available at: https://ir.nmu.org.ua/handle/123456789/150310.
28. Dubnytskyi, V. Yu., Krylenko, I. M., Fesenko, G. V. and Cherepnev, I. A. (2017), "The history of the development of means of eye protection for military personnel in combat conditions and modern requirements for controlling their impact resistance", *Weapons and military equipment systems*, No. 1(49), pp. 23-37, available at: https://www.hups.mil.gov.ua/periodic-app/article/17565.
29. Brovko, D. V. (2020), "Construction of a system for monitoring the reliability of elements of the built and constructed surface complex of mines based on entropy estimation", *Mining Bulletin: Scientific and Technical. coll.*, Kryvyi Rih, Issue 107, pp. 73–83, doi: https://doi.org/10.31721/2306-5435-2020-1-107-73-83.
30. Azarenkova, H. M., Zhuravel, T. M. and Mykhaylenko, R. M. (2009), "Enterprise finance: a study guide", Knowledge-Press, Kyiv, 299 p., available at: http://www.library.univ.kiev.ua/ukr/elcat/new/detail.php3?doc_id=1247203.
31. Prohonov, D. O. (2018), "Theoretical and informational evaluations of container distortions during the formation of steganograms", *Scientific and Technical Conference radioengineering fields, signals, devices and systems.* Conference Proceeding March 19-25, 2018, Kyiv, Ukraine, pp. 276-278, available at: http://ptmip.ipt.kpi.ua/list/progonov17.

32. Yehorshyn, O. O., Panova, N. V. and Polevych, V. V. (1955), *Regression analysis in examples and problems*, tutorial, Kharkiv State University of Economics, Kharkiv, 155 p.
33. Ulanova, E. S. and Zabelin, V. N. (1990), *Methods of correlation and regression analysis in agrometeorology*, Gidrometeoizdat, 207 p., available at: https://koha.lib.tsu.ru/cgi-bin/koha/opac-detail.pl?biblionumber=13364&shelfbrowse_itemnumber=40398.
34. Bondarenko V. N. (1970), "*Statistical solutions of some problems of geology*", NEDRA, Moscow, 244 p., available at: https://www.libex.ru/detail/book482854.html.
35. Chaddock, Robert Emmet (1925), *Principles and Methods of Statistics*, Hardcover, Houghton, Mifflin, 471 p., available at: https://books.google.com.sg/books/about/Principles_and_Methods_of_Statistics.html?id=-YxBTYcdnIoC&redir_esc=y.

СПИСОК ЛИТЕРАТУРИ

1. Мотало В. Аналіз шкал вимірювань. *Вимірювальна техніка та метрологія.* 2015. № 76. С. 21-35, URL: http://nbuv.gov.ua/UJRN/metrolog_2015_76_4
2. Maurice G. Kendall, Alan Stuart (1961). *The Advanced Theory of Statistics. Volume 2, Inference and Relationship*. Charles Griffin, London, 676 p.
3. Дубницкий В. Ю., Кобылин А. М., Кобылин О. А. Оценка нижней границы надёжности физически реализуемой системы в процессе её эксплуатации при произвольных законах распределения обобщённой нагрузки и прочности. *Системи обробки інформації*. 2018. № 1(152). С. 53-60. DOI: https://doi.org/10.30748/soi.2018.152.08.
4. Жлуктенко В. І., Наконечний С. І., Савіна С. С. Теорія ймовірностей і математична статистика. У 2-х ч. – Ч. ІІ. Математична статистика. Київ: КНЕУ, 2001. 336 с. URL: https://www.studmed.ru/zhluktenko-v-nakonechniy-s-savna-ss-teorya-ymovrnostey-matematichna-statistika-u-2-h-ch-ch-matematichna-statistika_3976c660ed4.html.
5. Kulldorff Gunnar (1961), Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples. Almqvist & Wiksell / John Wiley & Sons, Stokholm, 176 p. URL: https://www.amazon.com/Contributions-Estimation-Grouped-Partially-Samples/dp/B0010VDR26.
6. Джунь И. В. О числе градаций гистограмм ошибок астрономических наблюдений. *Кинематика и физика небесных тел.* 1993. №1, т. 9. С. 88-92. URL: https://www.mao.kiev.ua/biblio/jscans/kfnt/1993-09/kfnt-1993-09-1-11.pdf.
7. Джунь И. В. Математическая обработка астрономической и космической информации при негауссовых ошибках наблюдений: автореферат дис. на соиск. уч. степени докт. физ.-мат. наук: спец. 01.03.01 «Астрометрия и небесная механика». Киев, ГАО НАН Украины, 1992. 46 с. URL: https://issuu.com/blindguardian/docs/asd.
8. Паніотто В. І. , Максименко В. С., Харченко Н. М. Статаналіз соціологічних даних. Київ: КМ Академія, 2004. 270 с.
9. Цветков О. В. Энтропийный анализ данных в физике, биологии и технике. СПб.: СПбГЭТУ ЛЭТИ, 2015. 202 с. URL: https://www.researchgate.net/profile/Oleg-Tsvetkov/publication/331686300_entropijnyj_analiz_dannyh_v_fizike_biologii_i_tehnike/links/5c87f3afa6fdcc38174f8a14/entropijnyj-analiz-dannyh-v-fizike-biologii-i-tehnike.pdf.
10. Дубницкий В. Ю., Филатова Л. Д., Ходырев А. И. Устойчивость оценки энтропии гистограммы непрерывной случайной величины по отношению к изменению количества её интервалов. *Системи управління, навігації та зв'язку*. 2017, вип. 5(45), С. 42-46. URL: http://nbuv.gov.ua/UJRN/suntz_2017_5_12.
11. Agresti A. Categorical data analysis, John Wiley & Sons Inc., New York, 2002, 742 p. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/0471249688.
12. Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. Statistical Methods for Rates and Proportions. John Wiley & Sons, Inc. New York, 2003, 768 p. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/0471445428.
13. Nobuoki Eshima. *Statistical Data Analysis and Entropy*. Springer Nature Singapore Pte Ltd, Singapore, 2020, 498 p. URL: https://link.springer.com/book/10.1007/978-981-15-2552-0.
14. Graham J. G. Upton (1978), *The Analysis of Cross-tabulated Data*. J. Wiley, New York, 160 p. URL: https://www.amazon.com/Analysis-Cross-tabulated-Data-Graham-Upton/dp/0471996599.
15. Duncan Crammer. Advanced *Quantative Data Analysis*. Open University Press, Philadelphia, 2003. 272 p. URL: https://www.amazon.com/Advanced-Quantitative-Analysis-Understanding-Research/dp/0335200591.
16. Anne E. Magurran. *Ecological Diversity and its Measurement*. London, Sydney, CROOM HELM Royal Society University Research Fellow University, 1983, 184 p. URL: https://link.springer.com/book/10.1007/978-94-015-7358-0.
17. Margalef R. Information theory in ecology. *Gen. Syst.*, No 3, 1958. P. 36-71. URL: https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1134401.
18. Hutcheson K. A Test for Comparing Diversities Based on the Shannon Formula. Journal of Theoretical Biology. 29. 1970. P. 151- 154. URL: http://dx.doi.org/10.1016/0022-5193(70)90124-4.
19. Michalowicz, J.V., Nichols, J.M. and Bucholtz, F. *Handbook of differential entropy*, T.&F Gr, LLC, London, 2014. 241 p. URL: https://www.routledge.com/Handbook-of-Differential-Entropy/Michalowicz-Nichols-Bucholtz/p/book/9781138374799.
20. Linfoot E., Linfoot E. H. An Informational Measure of Correlation. *Information and Control*. Vol. 1, No. 1. 1957. P. 85-89. URL: https://www.sciencedirect.com/science/article/pii/S001999585790116X.
21. Vistelius A. B. The skew frequency distributions and the fundamental law of Geochemical processes. *The Journal of Geology*. Vol. 68, No. 1. 1960. P. 1-22. URL: https://www.jstor.org/stable/30058252.
22. Marian P. and Marian T. A. (2015), [Hellinger distance as a measure of Gaussian discord], *The Journal of Physics A: Math. Theor*., 48:11 (2015), 115301, 21 p., arXiv: 1408.4477. DOI: http://dx.doi.org/10.1088/1751-8113/48/11/115301.
23. Kullback S., Leibler R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, Vol. 22, No 1. 1951. pp. 79-86. DOI: http://dx.doi.org/10.1214/aoms/1177729694.
24. Дубницький В. Ю., Скорикова І. Г., Ходирєв О. І. Оптимальна апроксимація функції щільності розподілу за критерієм мінімуму втрати інформації. *Системи обробки інформації.* 2017. Вип. 4. С. 45-51. URL: https://www.hups.mil.gov.ua/periodic-app/article/17653.
25. Solomon Kullback. Information Theory and Statistics. Peter Smith, Gloucester, Mass. 1978. 399 p. URL: https://books.google.com.ua/books/about/Information_Theory_and_Statistics.html?id=XeRQAAAAMAAJ&redir_esc=y.
26. Гадецька С. В., Дубницький В. Ю., Кушнерук Ю. І., Ходирєв О. І. Спеціалізований програмний калькулятор для оцінки клінічної інформативності лабораторних тестів. *Сучасні інформаційні системи*. 2020. Т. 4, № 2. С. 80-84. DOI: https://doi.org/10.20998/2522-9052.2020.2.12.

27. Солошенко О. М. Дослідження відстані Кульбака-Лейблера у задачах моделювання у кредитному скорингу. *Развитие информационно-ресурсного обеспечения образования и науки в горно-металлургической отрасли и на транспорте 2014*: Днепропетровск, 2014. С. 328-333. URL: https://ir.nmu.org.ua/handle/123456789/150310.
28. Дубницький В. Ю, Криленко І. М., Фесенко Г. В., Черепньов І. А. Історія розвитку засобів захисту очей військовослужбовців в умовах бойових дій та сучасні вимоги до контролю їхньої стійкості до ударної дії. *Системи озброєння і військової техніки.* 2017, №1(49). С. 23-37. URL: https://www.hups.mil.gov.ua/periodic-app/article/17565.
29. Бровко Д. В. Построение системы мониторинга надежности элементов зданий и сооружений поверхностного комплекса шахт на базе оценки энтропии. Гірничий вісник, Кривий Ріг, 2020. Вип. 107. С. 73-83. DOI: https://doi.org/10.31721/2306-5435-2020-1-107-73-83.
30. Азаренкова Г. М., Журавель Т. М., Михайленко Р. М. Фінанси підприємств : навчальний посібник. Київ, Знання-Прес, 2009. 299 с. URL: http://www.library.univ.kiev.ua/ukr/elcat/new/detail.php3?doc_id=1247203.
31. Прогонов Д. О. Теоретико-інформаційні оцінки спотворень контейнерів при формуванні стеганограм. *Міжнародна науково-технічна конференція «Радіотехнічні поля, сигнали, апарати та системи». Київ, 19-25 березня 2018 р.* Київ, 2018. С. 276-278. URL: http://ptmip.ipt.kpi.ua/list/progonov17.
32. Єгоршин О. О., Панова Н. В., Полевич В. В. Регресійний аналіз у прикладах і задачах: навчальний посібник. Харків: ХНЕУ, 1999. 155 с.
33. Уланова Е. С., Забелин В. Н. Методы корреляционного и регрессионного анализа в агрометеорологии. Гидрометеоиздат, 1990. 207 с. URL: https://koha.lib.tsu.ru/cgi-bin/koha/opac-detail.pl?biblionumber=13364&shelfbrowse_itemnumber=40398.
34. Бондаренко В. Н. Статистические решения некоторых задач геологии. М.: НЕДРА, 1970. 244 с. URL: https://www.libex.ru/detail/book482854.html.
35. Chaddock Robert Emmet. *Principles and Methods of Statistics*. Hardcover, Houghton, Mifflin. 1925. 471 p. URL: https://books.google.com.sg/books/about/Principles_and_Methods_of_Statistics.html?id=-YxBTYcdnIoC&redir_esc=y.

Received (Надійшла) 22.01.2023
Accepted for publication (Прийнята до друку) 12.04.2023

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Гадецька Світлана Вікторівна – кандидат фізико-математичних наук, доцент, доцент кафедри вищої математики, Харківський національний автомобільно-дорожній університет, Харків, Україна;
Svitlana Gadetska – PhD in Physics and Mathematics, Associate Professor, Associate Professor of Department of Higher Mathematics of Kharkiv National Automobile and Highway University, Kharkiv, Ukraine;
e-mail: svgadetska@ukr.net; ORCID ID: https://orcid.org/0000-0002-9125-2363.

Дубницький Валерій Юрійович – кандидат технічних наук, старший науковий співробітник, старший науковий співробітник Харківського навчально-наукового інституту "Каразінський банківський інститут" Харківського національного університету ім. В. Н. Каразіна, Харків, Україна;
Valeriy Dubnitskiy – PhD in Engineering Senior Researcher Senior Researcher of "Karazin Banking Institute" of V.N. Karazin Kharkiv National University, Kharkiv, Ukraine;
e-mail: dubnitskiy@gmail.com; ORCID ID: https://orcid.org/0000-0003-1924-4104.

Кушнерук Юрій Іонович – кандидат технічних наук, доцент, доцент Інституту цивільної авіації Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна;
Yuri Kushneruk – Candidate of Technical Sciences, Associate Professor, Associate professor of Civil Aviation Institute of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: kyshneryk_ui@ukr.net; ORCID ID: https://orcid.org/0000-0001-5844-7137.

Ходирєв Олександр Іванович – старший викладач Харківського навчально-наукового інституту "Каразінський банківський інститут" Харківського національного університету ім. В. Н. Каразіна, Харків, Україна;
Alexander Khodyrev – Senior Lecturer of "Karazin Banking Institute" of V.N. Karazin Kharkiv National University, Kharkiv, Ukraine;
e-mail: khodyrevmjk3758@gmail.com; ORCID ID: https://orcid.org/0000-0001-9871-9440.

Excel-oriented calculator for calculating results of entropy analysis of data distributed by categories

Svitlana Gadetska, Valeriy Dubnitskiy, Yuri Kushneruk, Alexander Khodyrev

Abstract. The goal of the work. Development of EXCEL-oriented calculator for calculating the results of entropy analysis of data, which are distributed by categories. The subject of research is histograms of arbitrary distribution laws and conjugation tables 2×2. **Research methods:** Entropy and information analysis of histograms of arbitrary distribution laws and conjugation tables. **The obtained results.** It is proposed to use methods of entropy analysis for the analysis of data distributed by categories; information on the structure of the EXCEL-oriented calculator designed for this purpose is given. The calculator makes it possible to calculate entropy characteristics of histograms, namely: histogram entropy, histogram dispersion, histogram confidence intervals, diversity information index. The calculator performs a pairwise comparison of entropies of histograms using the Hutcheson method, determines Hellinger and Kullback-Leibler distances between histograms of arbitrary distribution laws and thus complements the chi-square criterion, determines the informational correlation coefficient. The correspondence between the Pearson correlation coefficient and the information correlation coefficient is established by the method of statistical modeling. For 2×2 conjugation tables, the calculator makes it possible to estimate the significance of the interaction between the row factor and the column factor. The calculator determines the values of conditional entropies for 2×2 conjugation tables. The proposed calculator fills the gaps in existing software products and can be used to process data distributed by categories using entropy analysis methods. It is shown that entropy methods of analysis are appropriate to use in cases where histograms determine arbitrary distribution laws.

Keywords: entropy; entropy analysis; information correlation coefficient; Hellinger distance; Kullback-Leibler distance.