УДК 004.421.2:519.2
DOI https://doi.org/10.32820/2074-8922-2019-63-134-142

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ В EXCEL ПРИ ВИКОНАННІ ІНДИВІДУАЛЬНО-РОЗРАХУНКОВОГО ЗАВДАННЯ В КУРСІ МАТЕМАТИЧНА СТАТИСТИКА

© Литвинов А.Л.
Харківський національний університет міського господарства ім. О.М. Бекетова

Інформація про автора:
Литвинов Анатолій Леонідович: ORCID: 0000-0001-7063-7814; litan@meta.ua; доктор технічних наук, професор кафедри прикладної математики і інформаційних технологій; Харківський національний університет міськогогосподарства ім. О. М. Бекетова; вул. Бажанова, 17, м. Харків, 61002, Україна.

У статті представлено результати з розробки інформаційної технології, яка автоматизує процес обробки масиву статистичних даних (вибірки) при виконанні індивідуально-розрахункового завдання (ІРЗ) у додатку Microsoft Office Excel. ІРЗ у рамках курсу «Математична статистика» відіграє особливу роль, дозволяє в рамках єдиного процесу комплексно використовувати отримані теоретичні знання. Розв'язок задач математичної статистики в рамках начальних програм для інженерних спеціальностей обумовлює значний обсяг обчислень, пов'язаний зі значним числом показників, складними алгоритмами їх розрахунку і графічною інтерпретацією результатів розв'язків. Традиційно в рамках ІРЗ студенту пропонується набір статистичних даних (вибірка) з невідомим законом розподілу. За цією вибіркою необхідно визначити оцінку для математичного сподівання даних вибірки, оцінку для дисперсії і середнього квадратичного відхилення, структурні середні, побудувати гістограму розподілу, висунути нульову гіпотезу про вид розподілу і за допомогою критерію χ^2 прийняти або відкинути висунуту гіпотезу. В ручному режимі ці розрахунки досить трудомісткі і монотонні, що вимагає використання комп'ютерної техніки при їх виконанні. Слід зазначити, що використання комп'ютерної техніки в навчальній літературі дисципліні «Математична статистика» приділяється вкрай мало уваги. В рамках курсу «Інформатика» розглядаються тільки окремі функції розрахунку статистичних характеристик, що ускладнює використання методів математичної статистики в навчальному процесі. Одним із найбільш доступних додатків для студентів є Excel. Тому однією з основних цілей нашої статті є розробка в середовищі Excel інформаційної технології послідовного розрахунку характеристик вибірки. Запропонована технологія складається з ряду етапів, для реалізації яких використовуються або вбудовані функції Excel, або програмовані вирази в клітинах електронної таблиці. Особлива увага приділена побудові гістограми розподілу, висунення гіпотези про вид розподілу і статистичної перевірки цієї гіпотези. Ефективність технології продемонстрована на прикладі.

Ключові слова: вибірка, інформаційна технологія, імовірність, критерій, математична статистика, параметр, показник, робочий лист, розподіл, Excel.

Литвинов А.Л. «Информационная технология в Excel при выполнении индивидуально-расчетного задания в курсе математическая статистика»

В статье представлены результаты по разработке информационной технологии, автоматизирующей процесс обработки массива статистических данных (выборки) при выполнении индивидуально-расчетного задания (ИРЗ) в приложении Microsoft Office Excel. ИРЗ в рамках курса «Математическая статистик» играет особую роль, позволяет в рамках единого процесса комплексно использовать полученные теоретические знания. Решение задач математической статистики в рамках учебных программ для инженерных специальностей обусловливает значительный объем вычислений, связанный со значительным числом показателей, сложными алгоритмами их расчета и графической интерпретацией результатов решений. Традиционно в рамках ИРЗ студенту предлагается набор статистических данных (выборка) с неизвестным законом распределения. По этой выборке необходимо определить оценку для математического ожидания данных выборки, оценку для дисперсии и среднего квадратического отклонения, структурные средние, построить гистограмму распределения, выдвинуть нулевую гипотезу о виде распределения и с помощью критерия χ^2 принять или отвергнуть выдвинутую гипотезу. В ручном режиме эти расчеты достаточно трудоемкие и монотонные, что требует использование компьютерной техники при их выполнении. Следует отметить, что использованию компьютерной техники в учебной литературе дисциплине «Математическая статистика» уделяется крайне мало внимания. В рамках курса «Информатика» рассматриваются только отдельные функции

расчета статистических характеристик, что затрудняет использование методов математической статистики в учебном процессе. Одним из самых доступных приложений для студентов является Excel. Поэтому одной из основных целей данной статьи является разработка в среде Excel информационной технологии последовательного расчета характеристик выборки. Предлагаемая технология состоит из ряда этапов, для реализации которых используются или встроенные функции Excel, или программируемые выражения в ячейках электронной таблицы. Особое внимание уделено построению гистограммы распределения, выдвижению гипотезы о виде распределения и статистической проверке этой гипотезы. Эффективность технологии продемонстрирована на примере.

Ключевые слова: выборка, информационная технология, вероятность, критерий, математическая статистика, параметр, показатель, рабочий лист, распределение, Excel.

A. Litvinov ″An information technology in Excel when executing an individual calculation task within the course Mathematical statistics″

The article presents the results of the process of elaborating an information technology which automatizes the processing of sequences of statistical data (samples) when executing an individually calculated task (ICT) in Microsoft Office Excel. An ICT plays a special role within the course “Mathematical Statistics”, and allows using the acquired theoretical knowledge comprehensively during a single process. Solving problems of mathematical statistics in the context of educational programs for engineering specialties requires a significant amount of computation, associated with a significant number of characteristics, complex algorithms for their calculation and graphical interpretation of the results of solutions. Traditionally, while working on an ICT, students are offered a set of statistical data (sample) with an unknown distribution law. From this sample, it is necessary to determine the estimate for the mathematical expectation of the sample data, the estimate for the dispersion and standard deviation, structural averages, build a histogram of the distribution, suggest a null hypothesis about the type of the distribution and use the $\chi 2$ criterion to accept or reject the hypothesis. In the manual mode, these calculations are rather time-consuming and monotonous, which requires the use of computer equipment when performing them. It should be noted that the use of computer technology in the educational literature with respect to the discipline “Mathematical Statistics” is given very little attention. Within the course “Computer Science”, only certain functions of calculating statistical characteristics are considered, which complicates the use of mathematical statistics in the educational process. Excel is one of the most available applications for students. Therefore, one of the main goals of this article is to develop an information technology in Excel for consistent calculation of sample characteristics. The proposed technology consists of a number of stages, for the implementation of which either built-in Excel functions or programmable expressions in spreadsheet cells are used. Particular attention is paid to building a distribution histogram, suggesting a hypothesis about the type of the distribution, and verifying statistically this hypothesis. The effectiveness of the technology is proved by the example provided.

Keywords: sample, information technology, probability, criterion, mathematical statistics, parameter, characteristics, worksheet, distribution, Excel.

Постановка проблеми. Однією з найважливіших складових частин навчальної програми з математичної статистики для інженерних спеціальностей є виконання індивідуально-розрахункового завдання (ІРЗ). ІРЗ посідає проміжне місце між практичними роботами, виконуваними в аудиторії і курсовою роботою (проектом). Особливо це стосується курсу «Математична статистика», для якого в рамках підготовки фахівців інженерного профілю курсова робота не передбачається. ІРЗ для курсу «Математична статистика» полягає в комплексній обробці масиву статистичних даних (вибірки) і пов’язано з великим обсягом обчислень за розрахунками великого числа показників, складними алгоритмами їх розрахунку і графічною інтерпретацією результатів розв’язків. Студенти можуть використовувати комп’ютерну техніку при проведенні розрахунків, але її застосування утруднене невеликою часткою часу, що приділяється використанню комп’ютерної техніки в рамках навчальної дисципліни. Слід зазначити, що використанню комп’ютерної техніки в навчальній літературі стосовно дисципліни «Математична статистика» приділяється вкрай мало уваги. В рамках курсу «Інформатика» розглядаються тільки окремі функції розрахунку статистичних характеристик, що ускладнює використання методів математичної статистики в навчальному процесі. У зв’язку з цим виникає проблема більш ширшого використання додатка Excel при вивченні дисципліни «Математична статистика».

Аналіз останніх досліджень і публікацій. З моменту своєї появи додаток Excel привернув увагу фахівців із різних галузей своєю функціональною завершеністю, широкими можливостями, легкістю освоєння. В [1] наведені основні відомості по роботі з Microsoft Excel. Значну увагу приділено форматуванню даних, роботі з елементами таблиці, редагування таблиць. У розділі «Розв'язування задач за допомогою формул» наведені відомості щодо введення формул в клітини таблиці, їх редагування і копіювання, використання вбудованих функцій. Що стосується математичної статистики, то в підручнику наведені лише початкові відомості і найпростіші функції з обробок вибірок (обчислення середнього, середньо квадратичного відхилення, мінімального і максимального значень). Немає загальної технології. У фундаментальній праці [2] розглянуто питання використання Excel в економіці. Книга орієнтована на підготовлених користувачів, що засвоїли ази Excel. Викладено технологію моделювання економічних систем і реалізацію моделей в Excel, зокрема фінансове моделювання, моделі прогнозування, моделі масового обслуговування. Особливу увагу приділено питанням оптимізації. Розглянуто задачі лінійного та нелінійного програмування, пошук екстремуму. Використанню Excel у математичній статистиці приділено мало уваги, лише в контексті з імовірнісним моделюванням. Стосовно використання Excel в математичній статистиці заслуговує уваги навчальний посібник [3]. У ньому послідовно викладається основний навчальний матеріал із математичної статистики з орієнтацією на студентів інженерного та економічного профілю. Так само, як і навчальні посібники [4, 5], він може використовуватися при виконанні індивідуально-розрахункового завдання в «ручному» режимі без використання обчислювальної техніки. Посібник складається з ряду розділів, кожен з яких присвячений розрахунку певним параметрам вибірки. Кожен розділ закінчується приведенням функцій Excel, за допомогою яких можна обчислювати відповідні параметри вибірки. Але в ньому немає послідовної технології обробки вибірок стосовно індивідуально-розрахункового завдання навчальної програми з математичної статистики.

Постановка завдань дослідження. Одним із найбільш доступних додатків для студентів є додаток Excel. Тому однією з основних цілей нашої статті є розробка в середовищі Excel інформаційної технології послідовного розрахунку характеристик вибірки. При виконанні індивідуально-розрахункового завдання студент стикається з низкою понять, розкриття змісту яких у підручниках не наводиться. Прикладом може служити число ступенів вільності, яке обчислюється за формулою $k=m-1-r$. Цим питанням також буде приділено увагу в нашій статті.

Виклад основного матеріалу. При освоєнні курсу «Математична статистика» однією з найбільш трудомістких завдань, які вирішують студенти, є комплеєю випадкової величини X значного обсягу. У зв'язку з великим обсягом обчислень, цю процедуру відносять до індивідуально-розрахункового завдання. Стандартна процедура включає такі етапи [4, 5]:

1. Визначення точкових оцінок вибірки. Вони повинні задовольняти ряду вимог, зокрема бути незміщеними, ефективними і обґрунтованими. Незміщеною оцінкою називають оцінку, математичне сподівання якої дорівнює оцінюваному параметру за будь-якого об'єму вибірки. Це гарантує від одержання систематичних помилок у більшу або меншу сторону. Зміщеною оцінкою називають оцінку, математичне сподівання якої не дорівнює оцінюваному. Ефективна оцінка – це така оцінка, що за конкретного об'єму вибірки має найменшу можливу дисперсію. Цим виключається можливість припуститися більшої помилки. Обґрунтована оцінка – це така оцінка, яка при необмеженому зростанні об'єму вибірки $(N \to \infty)$ прагне за ймовірністю до оцінюваного параметра. Якщо дисперсія незміщеної оцінки за $N \to \infty$ прагне до нуля, то така оцінка є обґрунтованою. Визначення точкових оцінок вибірки, що включає декілька етапів:

1.1. Визначення обсягу вибірки, тобто підрахунок кількості елементів вибірки – N.

1.2. Визначення оцінки середнього значення (математичного сподівання) випадкової величини, в якості якої зазвичай використовується вибіркове середнє $\overline{M}(X) = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Слід зауважити, що в залежності від об'єкту дослідження можуть використовуються такі види середніх: середня ступенева, середня гармонічна, середня квадратична, середня геометрична.

1.3. Знаходження оцінки для дисперсії, в якості якої зазвичай використовується вибіркова дисперсія і яка обчислюється за формулою $\overline{D}(X) = \overline{D} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$. На

відміну від вибіркового середнього, сума членів ряду поділяється на (N-1). Це гарантує, що оцінка дисперсії буде незміщеною [5].

Оскільки будь-яка оцінка є деяке наближення оцінюваної величини, то виникає питання про оцінку точності цього наближення. Поставимо задачу, обчисливши по вибірці оцінку для параметра Z – Z^*, указати для параметра Z такі дві границі $Z_н$ і $Z_в$, щоб істинне значення Z з деякою певною ймовірністю β, називаною довірчою, лежало в заданих границях, тобто щоб $P(Z_н < Z < Z_в) = \beta$. Відповідно інтервал ($Z_н$, $Z_в$) називається довірчим інтервалом, а кількості $Z_н$ і $Z_в$ – довірчими границями. Зазвичай довірчу ймовірність β вибирають близькою до одиниці (0,95; 0,99; 0,999). Величина $\Delta = (Z_в - Z_н)/2$ є точністю оцінки. Якщо випадкова величина розподілена нормально й відомо середньоквадратичне відхилення σ, довірчий інтервал для оцінки математичного сподівання m^* має границі

$$Z_н = \bar{x} - t_\beta \frac{\sigma}{\sqrt{N}} \text{ і } Z_в = \bar{x} + t_\beta \frac{\sigma}{\sqrt{N}}, \qquad (1)$$

де N – об'єм вибірки, а t_β знаходиться з рівності $\Phi(t_\beta) = \beta/2$ за таблицею значень функції Лапласа $\Phi(t)$. Зокрема, для β=0,90 t_β=1,6449, для β=0,95 t_β=1,9600, для β=0,99 t_β=2,575, для β=0,999 t_β=3,2905. З формули (1) випливає, що якщо потрібно забезпечити задану точність Δ оцінки математичного сподівання за заданої довірчої ймовірності, то мінімальний об'єм вибірки буде

$$N_{min} = t_\beta^2 \sigma^2 / \Delta^2 . \qquad (2)$$

Використовуючи формулу (2), студент може переконатися, чи є достатнім об'єм вибірки для проведення дослідження.

1.4. Знаходження оцінки для середнього квадратичного відхилення: $\bar{\sigma} = \sqrt{\bar{D}}$.

2. Визначення виду розподілу випадкової величини. Складається з таких етапів:

2.1. Визначення мінімального і максимального значення випадкової величини X, відповідно $X_{\min}$ і $X_{\max}$.

2.2. Знаходження діапазону зміни випадкової величини X в вибірці $H = X_{\min} - X_{\max}$.

2.3. Вибір числа розбиття діапазону зміни випадкової величини X на m часткових інтервалів. Чим більше обсяг вибірки, тим більше можна взяти число m і тим точніше будуть результати досліджень. Для вибору числа m можна використовувати наближену формулу Стерджесса: $m = 1 + 1{,}44 \ln N$ [6, 7].

2.4. За формулою $h = H / m$ розраховується довжина часткових напівінтервалів. Доцільно її скорегувати, щоб була «зручною» для розрахунків. Перший і останній напівінтервалів також можна скорегувати для зручності обчислень.

2.5. Підрахунок кількості влучень випадкової величини X у кожен частковий напівінтервал (частота) - n_i. При ручному рахунку це найбільш трудомісткий і в той самий час найбільш відповідальний етап виконання індивідуально-розрахункового завдання, від якого багато в чому залежить достовірність отриманих результатів. Можна попередньо впорядкувати масив даних вибірки, після чого розрахунок частот n_i не складе труднощів. Недолік - висока трудомісткість. Можна запропонувати таку технологію.

Формуємо таблицю емпіричних частот такого виду (див. табл. 1).

Таблиця 1

Таблиця емпіричних частот

№	Діапазон напівінтервала	варіанти	n_i
1	$[x_0 - x_1]$		n_1
2	$(x_1 - x_2]$		n_2
3	$(x_2 - x_3]$		n_3
…	……..	……	…
m	$(x_{m-1} - x_m]$		n_m

Послідовно аналізуючи дані вибірки, розносимо варіанти в клітинки стовпця «варіанти». Після чого підраховуємо кількість варіантів у кожній клітинці і заповнюємо стовпець «n_i».

2.6. Розрахунок щільності відносних частот для кожного напівінтервала за формулою $\rho_i = n_i / (Nh)$.

2.7. Побудова гістограми відносних частот.

2.8. За формою гістограми відносних частот висувається так звана нульова гіпотеза про вид розподілу. Вибір закону розподілу зазвичай здійснюється за графіком щільності розподілу. Їх можна знайти в довідниках [8]. Зазвичай для індивідуально-розрахункового завдання підбираються вибірки, закон розподілу

яких підпорядковується або нормальному, або експоненціальному, або рівномірному закону.

2.9. Визначення параметрів розподілу. Цей пункт можна виконувати або методом моментів або методом найбільшої правдоподібності [9]. Слід зауважити, для виконання індивідуально-розрахункового завдання студентами інженерного профілю достатньо використовувати метод моментів, який полягає в прирівнюванні початкових або центральних моментів (математичного сподівання, дисперсії тощо) теоретичного розподілу, що є функціями параметрів розподілу, до відповідних моментів, обчислених з вибірки, і розв'язання одного або декількох рівнянь щодо параметрів розподілу.

3. Перевірка нульової гіпотези про вид обраного розподілу.

3.1. Перш за все це вибір критерію згоди *K*. Зазвичай використовуються одномірний критерій згоди Пірсона хі-квадрат (χ^2), який обчислюється за формулою

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - Np_i)^2}{Np_i} = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'}, \qquad (3)$$

де у чисельнику стоїть різниця між експериментальною і теоретичною частотами для кожного напівінтервала. Вираз (3) розподілено по закону χ^2 («хі квадрат», який має щільність розподілу

$$f(x) = \begin{cases} 0 & \text{при } x \le 0 \\ \dfrac{1}{2^{k/2}\Gamma(k/2)} exp(-x/2)x^{0,5k-1} & \end{cases}, \qquad (4)$$

де $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ – гама-функція.

Параметр *k* у виразу (2) це так званий ступень свободи (degree of freedom), який увів видатний дослідник Рональд Фішер і який означає число незалежних додатків у сумі з урахуванням внутрішніх зв'язків, які віднімаються від числа часткових напівінтервалів. До цих внутрішніх зв'язків належить, по-перше, кількість даних у виборці, а, по-друге, кількість параметрів вибраного закону розподілу – *r*. Тобто число ступенів свободи *k* обчислюється за формулою *k*=*m*-1-*r*. Наприклад, якщо використовується нормальний закон розподілу, то він характеризується двома параметрами, відповідно *b* і σ. *b* – математичне сподівання, σ – середньо квадратичне відхилення. Звідкіля *r* дорівнює двом. Для критерію χ^2 розраховується межа правобічної критичної області $k_{кр}$ при заданому рівні значущості α – ймовірності відхилення нульової гіпотези, і числа ступенів свободи. При цьому ймовірність того, що критерій *K* набуде значення більше за $k_{кр}$, буде дорівнювати прийнятому рівню значущості $-P(K > k_{кр}) = \alpha$. Значення $k_{кр}$ при заданих рівні значущості α і числа ступенів вільності за звичай приводяться в додатках більшості навчальних посібниках з математичної статистики. Слід зауважити, що для обчислення $k_{кр}$ і других характеристик, пов'язаних із розподілом χ^2 в Excel є ряд функцій, зокрема =ХИ2.ОБР.ПХ(рівень_значущості; число_ступенів_свободи).

3.2. За вибіркою розраховується спостережуване значення критерію згоди – $K_{спост}$. Якщо для використання вибрано критерій згоди Пірсона хі-квадрат (χ^2), то він розраховується за формулою:

$$K_{спост} = \chi^2_{спост} = \sum_{i=1}^{m} \frac{(n_i - Np_i)^2}{Np_i} = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'}. \qquad (5)$$

3.3. Якщо $K_{спост} > K_{кр}$, то висунуту гіпотезу відкидаємо і пробуємо підібрати інший розподіл. У ряді випадків необхідно збільшити обсяг вибірки. Якщо $K_{спост} < K_{кр}$, то висунуту гіпотезу приймаємо і розраховуємо так званий *p*-level – максимальна ймовірність при заданому $K_{спост}$, при якій ще справедлива нульова гіпотеза. Він дозволяє оцінити ступінь впевненості в обраній гіпотезі про вид розподілу. Для розрахунку *p*-level можна використовувати таблицю критичних точок розподілу. Для цього потрібно в таблиці для критичних точок розподілу знайти найближче до $\chi^2_{спост}$ значення $\chi^2_{спост}$ ($\chi^2 > \chi^2_{спост}$) для заданої кількості ступенів свободи и подивитись відповідній йому рівень значущості α. У додатку Microsoft Excel, починаючи з версії 2010, є спеціальна функція для обчислення ймовірності *p*-level: =ХИ2.РАСП.ПХ(X;Степени_свободы), де **X** – досліджуваний $\chi^2_{спост}$, Степени_свободы – кількість ступенів свободи *k*=*m*-1-*r*.

4. Крім перерахованих видів середніх, у індивідуально-розрахунковому завданні можуть обчислювати так звані структурні середні, які не зв'язані зі значеннями варіант, розташованих на кінцях розподілу, а пов'язані з рядом частот. До них належать медіана й мода. Медіана – M_e^* статистичного розподілу вибірки – це варіанта, яка поділяє варіаційний ряд, утворений після первинного сортування простого статистичного

ряду, на два рівні за кількістю варіант частини. Якщо кількість варіант непарне, тобто $N = 2k+1$, тоді $M_e^* = x_{k+1}$. Якщо кількість варіант парне, тобто $N = 2k$, тоді медіана дорівнює середньому арифметичному "серединної" (медіанної) пари варіант: $M_e^* = (x_k + x_{k+1})/2$. Мода – M_0^* – статистичного розподілу вибірки це варіанта, яка має найбільшу частоту. Мода визначається безпосередньо за даними статистичного розподілу. Вона широко використається в комерційній діяльності, у соціологічних дослідженнях, коли вивчається ринковий попит, при реєстрації цін, установленні рейтингу популярності лідерів або товарів. Варіаційний розмах – різниця між крайніми (найбільшою і найменшою) варіантами: $R = x_{\max} - x_{\min}$.

З аналізу технології виконання індивідуально-розрахункового завдання видно, його виконання дуже трудомістке. Пропонується така інформаційна технологія автоматизації вище наведених етапів із використанням програми MS Excel.

1. Занесення масиву даних, що входять до вибірки в робочий лист MS Excel.

2. Визначення об'єму вибірки за допомогою функції СЧЁТ( ) з категорії «Статистичні».

3. Побудова гістограми частот. Для цього використовуємо інструмент аналізу «Гістограма» з надбудови «Пакет аналізу». На вкладці «Дані» ця надбудова, якщо вона встановлена, відображається в групі «Аналіз» як «Аналіз даних». Якщо надбудова «Пакет аналізу» не встановлена, то її необхідно встановити, послідовно відпрацьовуючи «Файл» → «Параметри» → «Надбудови» → Управління «Надбудови Excel» → «Перейти» → установка галочки на «Пакет аналіз» → ОК.

Наступним кроком буде формування числа часткових інтервалів, їх меж за зростанням і занесення масиву границь часткових інтервалів, починаючи з крайнього лівого в лист Excel. Для обчислення число часткових напівінтервалів за формулою Стерджесса використовуємо комбінацію функцій Excel : =ОКРУГЛВВЕРХ(1+1,44*LN();0).

За допомогою команд =МІН( ) і =МАКС( ) із категорії «Статистичні» знаходимо мінімальне і максимальне значення варіант у вибірці, відповідно $x_0 = X_{min}$ і $x_m = X_{max}$. За формулою $H = X_{max} - X_{min}$ розраховуємо діапазон зміни випадкової величини X і по формулі $h = H / m$ знаходимо довжину часткового інтервалу. У загальному випадку значення x_0 може коригуватися в меншу сторону, а x_m - в більшу сторону. Розраховуємо границі часткових напівінтервалів:

$$x_1 = x_0 + h, x_2 = x_1 + h = x_0 + 2h, ..., x_i = x_0 + ih, i < m.$$

Нехай $x_1, x_2, x_3, ..., x_{m-1}$ – точки, які розбивають весь діапазон зміни випадкової величини X на m часткових інтервалів. В інструменті «Гістограма» їм будуть відповідати такі напівінтервали:

$$(-\infty; x_1], (x_1; x_2], (x_2; x_3], ..., (x_{m-1}, x_m], (x_m, +\infty),$$

і для цих часткових напівінтервалів будуть підраховані частоти потрапляння значень випадкової величини X – n_i, $i = 1, 2, ...$ Якщо в масиві вихідних даних знайдуться дані, які перевершують x_m, то їхня кількість буде відображено в кишені під ім'ям «Ще». У лист Excel заносимо масив чисел $x_1, x_2, x_3, ..., x_m$, який буде інтерпретуватися в «Гістограма» як параметр «Інтервал кишень».

По ланцюгу «Дані → Аналіз_данних → Гістограма» викликаємо додаток побудови гістограми і у вікні «Гістограма» заповнюємо відповідні параметри («Вхідний інтервал», «Інтервал кишень»), вказуємо на новий робочий лист і задаємо вивід графіка. В результаті отримаємо таблицю частот і гістограму частот. По виду гістограми висуваємо гіпотезу про вигляд закону розподілу, якому підпорядковуються дані вибірки. Як правило, для індивідуально-розрахункового завдання підбираються вибірки, закон розподілу яких підпорядковується або нормальному, або експоненціальному, або рівномірному закону.

Використовуючи метод моментів, визначаємо параметри закону розподілу і вираз для функції розподілу. Якщо отримали експоненціальний розподіл (щільність розподілу $f(t)=\exp(-\lambda t)$, $t>0$), то він має один параметр λ, який дорівнює 1/m, де m – математичне сподівання, у якості якого можна використовувати оцінку для середнього вибірки [9]. Якщо отримали нормальний закон розподілу, то він характеризується двома параметрами, відповідно b і σ. b – математичне сподівання, σ – середньо квадратичне відхилення, у якості яких можна взяти відповідні оцінки. Якщо отримали рівномірний закон (щільність розподілу $f(t)=1/(b-a)$), то для обчислення його параметрів можна використовувати такі формули: $b = \bar{x} + \sqrt{3}\bar{\sigma}$, $a = \bar{x} - \sqrt{3}\bar{\sigma}$.

4. Розрахунок $\chi^2_{спост}$.

4.1. Видаляємо гістограму, додаємо рядок нижче першого і заносимо границю лівої кишені (лівого півінтервалу).

4.2. Праворуч від стовпчика «Частота» організуємо стовпець «Теоретична частота», де для кожної кишені розраховуємо теоретичну частоту n_i' за формулою $N \cdot (F(x_{прав}) - F(x_{лів}))$. F(x) – теоретична функція розподілу. Для контролю підсумовуємо всі значення, вони мають бути близькими до N.

4.3. Праворуч від стовпчика «Теоретична частота» організуємо стовпець із частковими значеннями критерію $\chi^2_{спост}$, які обчислюються за формулою $(n_i - n_i')^2 / n_i'$.

4.4. Підсумувавши значення цього стовпця, знайдемо значення $\chi^2_{спост}$.

5. Задаємо значення α і число ступенів вільності. Скориставшись функцією =ХИ2.ОБР.ПХ(α;Степени_свободы), знаходимо χ^2_{kp} і порівнюємо його з $\chi^2_{спост}$. Якщо $\chi^2_{спост} < \chi^2_{kp}$, то приймаємо нульову гіпотезу і за допомогою функції =ХИ2.РАСП.ПХ($\chi^2_{спост}$;Степени_свободы) заходим ймовірність p-level. Рівень p-level, який його задовольняє, дослідник вибирає самостійно

6. Лист доповнюємо діаграмою у вигляді гістограм частот і теоретичних частот.

Доцільно для успішного використання запропонованої технології, щоб довжини всіх карманів, тобто півінтервалів, на які розбивається весь діапазон зміни випадкової величини, були однакові.

Проведемо демонстрацію запропонованої технології на даних вибірки, що містить інтервали часу між сусідніми запитами, які надходять до мережевого принтера – t_i (табл. 2):

Таблиця 2

Дані вибірки до прикладу

	8,97	70,5	28	5,27	18,3	17,6	3,9	17,5	21,3	59,7	1,73
22,4	63,10	7,82	34,2	7,39	28,3	86,2	10,2	3,3	7,7	13,9	0,701
63,5	2,38	65,1	24,2	0,116	38,3	37,3	30,8	74,7	23,4	59,3	20,1
25,1	76,6	10,9	20,1	1	25,9	5,16	67	33,1	21,1	11	43,5
26,9	27,7	31,5	0,3	9	20,6	14,3	49,3	6,85	2,54	62,4	15,5
44,7	60,4	65,8	17,5	4,58	119	6,04	8,74	18,7	1,61	40	26,8
35,8	26,6	16	38,5	0,39	110	17,9	1,89	27,9	5,24	6,62	25,6
1,41	47,4	13,1	3,25	31,6	58,1	21	39,7	43,9	27,6	24	18
7,78	21,7	1,93	26,7								

Заносимо дані вибірки у лист Excel (див. рис. 1).

	A	B	C	D	E	F	G	H	I	J	K	L
1					ВИБІРКА							
2	7,35	8,97	70,5	28	5,27	18,3	17,6	3,9	17,5	21,3	59,7	1,73
3	22,4	63,1	7,82	34,2	7,39	28,3	86,2	10,2	3,3	7,7	13,9	0,701
4	63,5	2,38	65,1	24,2	0,12	38,3	37,3	30,8	74,7	23,4	59,3	20,1
5	25,1	76,6	10,9	20,1	1	25,9	5,16	67	33,1	21,1	11	43,5
6	26,9	27,7	31,5	0,3	9	20,6	14,3	49,3	6,85	2,54	62,4	15,5
7	44,7	60,4	65,8	17,5	4,58	119	6,04	8,74	18,7	1,61	40	26,8
8	35,8	26,6	16	38,5	0,39	110	17,9	1,89	27,9	5,24	6,62	25,6
9	1,41	47,4	13,1	3,25	31,6	58,1	21	39,7	43,9	27,6	24	18
10	7,78	21,7	1,93	26,7								
11									закон	розподілу		
12	N=	100		H=	120	$\bar{t}$ =	27,054		експоненціальний			
13	m=	8		h=	15	σ=	24,399		$\lambda = 1/\bar{t} = 0{,}037$			
14	Xmin=	0										
15	Xmax=	120		Кармани:	15	30	45	60	75	90	105	120

Рис. 1 – Робочий лист Excel із розрахунками до прикладу

За допомогою функції =СЧЕТ( ) знаходимо кількість даних у вибірці – 100. За допомогою функцій =МИН( ) і =МАКС( ) знаходимо мінімальний і максимальний елементи вибірки. Відповідно, min = 0,116 ≈ 0, max = 119. Щоб уникнути ручного округлення, доцільно використовувати функції: =ОКРВНИЗ(МИН( );1) и =ОКРВВЕРХ(МАКС( );10)

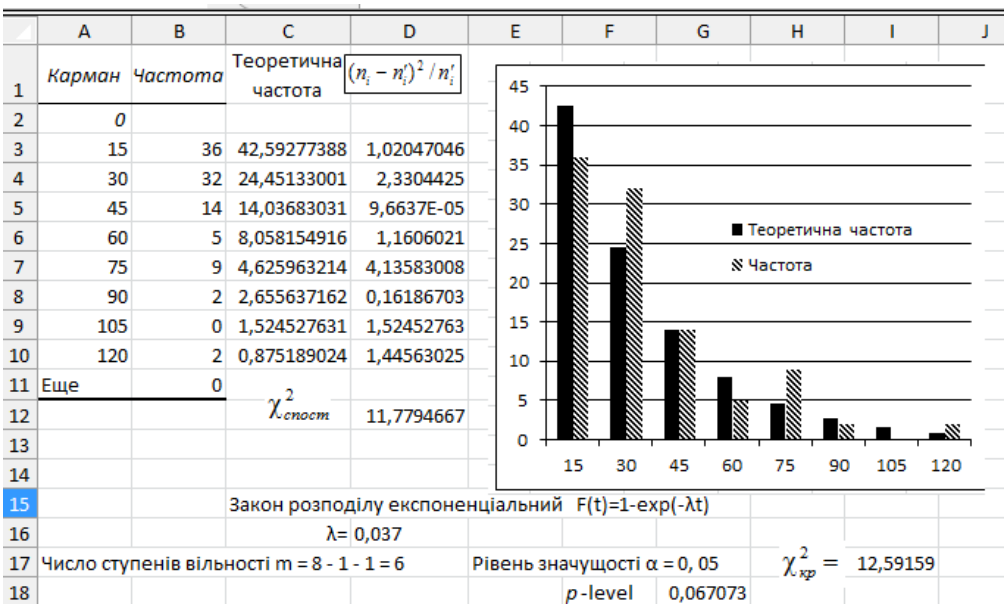
Знаходимо діапазон даних у вибірці: $H = max - min$; за необхідностю округляємо в більшу сторону, для чого використовує функцію =ОКРВВЕРХ(МАКС - MIN);10). Отримали $H = 120$.

Використовуючи комбінацію функцій, у яку входить формула Стерджесса =ОКРУГЛВВЕРХ(1+1,44*LN( );0), вибираємо кількість часткових пів інтервалів, $m = 8$.

Обчислюємо довжину кожного півінтервала, на які розбивається діапазон зміни даних: $h = H / m = 15$.

За формулою $t_i = t_{i-1} + h,\ i = 1,2,\ ...,8;\ t_0 = MIN$ формуємо границі карманів і розміщаємо їх у послідовні клітинки (див. рис.1).

По ланцюгу «Дані → Аналіз_данних → Гістограма» викликаємо додаток побудови гістограми і у вікні «Гістограма» заповнюємо відповідні параметри («Вхідний інтервал», «Інтервал кишень»), вказуємо на новий робочий лист і задаємо вивід графіка. У результаті отримаємо таблицю частот і гістограму частот (див. рис.2).



	A	B	C	D
1	Карман	Частота	Теоретична частота	$(n_i - n_i')^2 / n_i'$
2	0			
3	15	36	42,59277388	1,02047046
4	30	32	24,45133001	2,3304425
5	45	14	14,03683031	9,6637E-05
6	60	5	8,058154916	1,1606021
7	75	9	4,625963214	4,13583008
8	90	2	2,655637162	0,16186703
9	105	0	1,524527631	1,52452763
10	120	2	0,875189024	1,44563025
11	Еще	0		
12			$\chi^2_{спост}$	11,7794667
13				
14				
15			Закон розподілу експоненціальний F(t)=1-exp(-λt)	
16			λ= 0,037	
17	Число ступенів вільності m = 8 - 1 - 1 = 6		Рівень значущості α = 0, 05	$\chi^2_{кр}$ = 12,59159
18			p-level	0,067073

Рис. 2 – Лист Excel із результатами обробки вибірки

За виглядом гістограми частот висуваємо гіпотезу – інтервали часу між моментами надходження запитів до мережевих принтера на друк підпорядковуються експоненціальному закону з функцією розподілу $F(t) = 1 - exp(-\lambda t)$.

Знаходимо параметр λ. Для експоненціального закону $\lambda = 1/\bar{t}$, де $\bar{t}$ – вибіркова середня. Використовуючи комбінацію функцій =ОКРУГЛТ(1/СРЗНАЧ( );0,0001), знаходимо, що $\lambda = 0{,}0370$.

При рівні значущості $\alpha = 0{,}05$ перевіряємо справедливість висунутої гіпотези.

Видаляємо гістограму, додаємо рядок нижче першого і заносимо границю лівої кишені (лівого півінтервалу) – 0.

Праворуч від стовпчика «Частота» організуємо стовпець «Теоретична частота», де для кожної кишені розраховуємо теоретичну частоту n_i' за формулою

$$= 100 * (\text{EXP}(-0{,}037 * t_{i-1}) - \text{EXP}(-0{,}037 * t_i)).$$

Праворуч від стовпчика «Теоретична частота» організуємо стовпець із частковими значеннями критерію $\chi^2_{спост}$, які обчислюються за формулою $(n_i - n_i')^2 / n_i'$. Згідно з рис. 2, для цього використовується формули $= (B_i - C_i)\text{^}2 / C_i$. Підсумувавши часткові значення, знаходимо значення критерію хі квадрат, який спостерігається – $\chi^2_{спост} = 11{,}779$.

За функцією =ХИ2.ОБР.ПХ(0,05;6), де 0,05 – рівень значущості, а 6=8-1-1 – число стенів свободи для вибраного числа півінтервалів і експоненціальному закону розподілу, знаходимо критичне значення критерію хі-квадрат – $\chi^2_{кр} = 12{,}591$.

Оскільки $\chi^2_{спост} < \chi^2_{кр}$, то можна прийняти гіпотезу про експоненціальний (показниковий) розподіл випадкової величини *T* – інтервали часу між моментами надходження запитів до мережевого принтера на обслуговування.

За функцією =ХИ2.РАСП($\chi^2_{спост}$, число_ступенів_свободи) розрахуємо *p*-level – максимальна ймовірність при заданому $\chi^2_{спост}$, при якій ще виконується нульова гіпотеза. Отримали: level = 0,0670. Так як *p*-level > α, то це також підтверджує нульову гіпотезу – розподіл випадкової величини *T* – інтервали часу між моментами надходження запитів до мережевого принтера на обслуговування підкоряються експоненціальному розподілу з параметром λ = 0,0370. На завершення будуємо сумісну гістограму теоретичної і експериментальних частот (див. рис. 2). Візуально вона підтверджує висунуту гіпотезу, про експоненціальний закон розподілу даних вибірки.

Висновки і перспективи подальших досліджень. Апробація запропонованої інформаційної технології при викладанні курсу «Математична статистика» для інженерних та економічних спеціальностей дозволила значно прискорити проведення розрахунків при виконанні практичних занять та індивідуально-розрахункового завдання, подолати розрив між досить складним теоретичним матеріалом і його практичним застосуванням. Слід зазначити, що розроблена інформаційна технологія стосується одновимірного аналізу. Багатовимірний статистичний аналіз відображений в Excel тільки декількома функціями (PEARSON (массив1; массів2), КВПІРСОН, КОВАРІАЦІЯ.В (Г), КОРРЕЛ) з обмеженими можливостями. Значні труднощі виникають при обробці двовимірних кореляційних таблиць. Подальші дослідження повинні бути спрямовані на розробку інформаційних технологій обробки багатовимірних вибірок у рамках додатка Excel.

Список використаних джерел

1. Завадський І. О. Microsoft Excel у профільному навчанні : навч. посіб. / І. О. Завадський, А. П. Забарна. – Київ : Вид. група ВНУ, 2011. – 272 с.

2. Экономическое моделирование в Microsoft Excel / Д. Мур, Л. Р. Уэдерфорд, Г. Эппен [и др.]. – М. : Вильямс, 2004. – 1024 с.

3. Воскобойников Ю. Е. Математическая статистика (с примерами в Excel.) : учеб. пособие / Ю. Е. Воскобойников. Е. И. Тимошенко. – Новосибирск : Новосиб. гос. архитектур.-строит. ун-т (Сибстрин), 2006. – 152 с.

4. Волощенко А. Б. Теорія ймовірностей та математична статистика : навч.-метод. посібник для самост. вивч. дисц. / А. Б. Волощенко, І. А. Джалладова. – Київ : КНЕУ, 2003. – 256 с.

5. Математична статистика: навчальний посібник / Є. О. Лебєдєв, Г. В. Лівійська, І. В. Розора, [та ін.]. – Київ : ВПЦ "Київський університет", 2016. – 160 с.

6. Математическая статистика: практикум / О. Б. Тарасова, Е. В. Шайкина, А. Е. Шибалкин [и др.]. – М. : Изд-во РГАУ-МСХА имени К. А. Тимирязева, 2014. –136 с.

7. Борисова Е. В. Прикладные статистические модели и методы в социологии : учебное пособие / Е. Б. Борисова. – М. : АНАЛИТИКА РОДИС, 2016. – 254 с.

8. Хастингс Н. Справочник по статистическим распределениям / Н. Хастингс, Дж. Пикок. – М. : Статистика, 1980. – 95 с.

9. Гмурман В. Е. Теория вероятностей и математическая статистика / В. Е. Гмурман. – М. : Высш. шк., 2002. – 479 с.

Referenses

1. Zavadskyi, IO & Zabarna, AP 2011, *Microsoft Excel u profilnomu navchanni*, [Microsoft Excel in profile training] Vydavnycha hrupa Skhidnoukrainskoho natsionalnoho universytetu, Kyiv.

2. Mur, D, Uederford, LR, Eppen, G, Guld, F & Shmidt, Ch 2004, Jekonomicheskoe modelirovanie v Microsoft Excel, [Economic Modeling in Microsoft Excel] Viljams, Moskva.

3. Voskobojnikov, JuE & Timoshenko, EI 2006, *Matematicheskaja statistika (s primerami v Excel)*, [Mathematical statistics (with examples in Excel.)] Novosibirskij gosudarstvennyj arhitekturno-stroitelnyj universitet (Sibstrin), Novosibirsk.

4. Voloshchenko, AB & Dzhalladova, IA 2003, *Teoriia ymovirnostei ta matematychna statystyka*, Kyivskyi natsionalnyi ekonomichnyi universytet, Kyiv.

5. Lebiediev, YeO, Liviiska, HV, Rozora IV & Sharapov, MM 2016, *Matematychna statystyka*, [Mathematical statistics: a textbook] Vydavnycho-polihrafichnyi tsentr Kyivskyi universytet, Kyiv.

6. Tarasova, OB, Shajkina, EV, Shibalkin, AE & Kagirova, MV 2014, *Matematicheskaja statistika*, [Mathematical Statistics: Workshop] Izdatelstvo Rossijskogo gosudarstvennogo agrarnogo universiteta - MSHA imeni K. A. Timirjazeva, Moskva.

7. Borisova, EV 2016, *Prikladnye statisticheskie modeli i metody v sociologii*, [Applied statistical models and methods in sociology: a training manual] ANALITIKA RODIS, Moskva.

8. Hastings, N & Pikok, Dzh 1980, *Spravochnik po statisticheskim raspredelenijam*, [Statistical Distribution Handbook] Statistika, Moskva.

9. Gmurman, VE 2002, *Teorija verojatnostej i matematicheskaja statistika*, [Theory of Probability and Mathematical Statistics] Vysshaja shkola, Moskva.

Стаття надійшла до редакції 30.07.2019 р.