Міністерство освіти і науки України
Черкаський національний університет
імені Богдана Хмельницького

ГАЛИНА ЛУЦЕНКО

АНАЛІЗ І ВІЗУАЛІЗАЦІЯ ДАНИХ

Навчально-методичний посібник для виконання лабораторних робіт

ЧЕРКАСИ
2024

УДК 378.1
Л86

Автор і укладач: Луценко Г.В., д.п.н., доцент

Рецензенти: Трифонова О. М., д.п.н., професор,
Подолян О.М., к.ф.-м.н., доцент

Л86 Луценко Г.В. (2024). Аналіз і візуалізація даних: навчально-методичний посібник для студентів закладів вищої освіти спеціальності 014 Середня освіта (014.09 Інформатика). Черкаси: ЧНУ ім. Б. Хмельницького. – 90 с.

У навчально-методичному посібнику подано методичні вказівки для виконання лабораторних робіт з дисципліни «Аналіз і візуалізація даних» для здобувачів вищої освіти першого (бакалаврського) рівня за спеціальністю 014 Середня освіта (014.09 Інформатика) денної форми навчання.

Затверджено на засіданні кафедри автоматизації та комп’ютерно-інтегрованих технологій, протокол №9 від 15 березня 2024 р.

Затверджено на засіданні вченої ради Черкаського національного університету імені Богдана Хмельницького, протокол №12 від 20 червня 2024 р.

ЗМІСТ

ВСТУП

Відповідно до освітньо-професійної програми «Інформатика» підготовки здобувачів першого (бакалаврського) рівня вищої освіти за спеціальністю 014 Середня освіта (014.09 Інформатика) до переліку обов’язкових компонентів ОПП входить дисципліна «Аналіз і візуалізація даних».

Аналіз даних є процесом перетворення даних, отриманих різними способами, у корисну інформацію. Аналіз даних включає виявлення патернів для досліджуваних наборів значень, взаємозв'язків та трендів у даних, що допомагає приймати обґрунтовані рішення та робити прогнози. Візуалізація даних – це процес представлення даних і їх поведнки у вигляді графіків, діаграм, карт, що допомагає презентувати інформацію у доступнішій та зрозумілішій для аудиторії формі. Візуалізація дозволяє відобразити складні зв'язки та закономірності даних, шляхом графічного зображення, що також полегшує прийняття рішень.

Дисципліна «Аналіз і візуалізація даних» спрямована на формування у майбутніх учителів інформатики здатності орієнтуватися в інформаційному просторі, шукати, обробляти, організовувати, візуалізувати й критично оцінювати інформацію, оперувати нею у професійній діяльності. Ключовим інструментарієм цього курсу є теорія, принципи і методи математичної статистики, використання яких забезпечує якісне проведення комп’ютерних експериментів у частині очищення, обробки й представлення даних.

Силабус дисципліни «Аналіз і візуалізація даних» передбачає проведення лабораторних занять, у ході яких студенти детально опрацьовують теоретичних матеріал і набувають практичних навичок із застосування статистичних методів при обробці експериментальних даних. У навчально-методичному посібнику детально висвітлюються практичні аспекти використання цифрових інструментів для аналізу й візуалізації даних. Він містить необхідні теоретичні матеріали, типові приклади та завдання для кожного з розділів, які охоплюють первинну статистичну обробку результатів спостережень та перевірку статистичних гіпотез щодо законів розподілу, вибірок та числових характеристик.

ОПИС КУРСУ

Метою дисципліни «Аналіз і візуалізація даних» є забезпечення формування теоретичних знань й навичок практичного застосування у професійній діяльності понятійного апарату, принципів, методів та програмного забезпечення аналізу та візуалізації даних. Вивчення дисципліни "Аналіз і візуалізація даних" дозволить студентам краще зрозуміти специфіку роботи з даними різної природи. Під час вивчення курсу студенти навчатимуться працювати з наборами даних, отриманими з відкритих джерел й методами формування статистичних вибірок, форматувати дані, відповідно до встановлених критерії, обирати оптимальні методи статистичної обробки, оцінювати й критично аналізувати отримані результати, візуалізувати їх.

Аналіз даних як дисципліна оперує поняттями математичної статистики, вимагає розуміння предметної області дослідження та засад програмування. У курсі «Аналіз і візуалізація даних» розкриваються предмет, методи та базові категорії математичної статистики; вивчаються міри центральної тенденції та мінливості вибірок, статистичне оцінювання, перевірка статистичних гіпотез з використанням параметричних і непараметричних критеріїв, кореляційний і регресійний аналіз. Розглядаються технологічні прийоми і способи комп'ютерної реалізації аналізу й візуалізації даних на базі табличного процесора Google Таблиці. Також у курсі розглядаються сучасні підходи до отримання, очищення, структурування та збереження даних за фаховою спрямованістю.

Завданнями курсу визначено:

1. Аналіз розподілу даних: Використовуючи набір даних, студенти повинні побудувати гістограму та визначити тип розподілу (нормальний, рівномірний тощо). Додатково, обчислити середнє значення, медіану та стандартне відхилення.

2. Оцінка кореляції: На основі реального набору даних, студенти повинні провести кореляційний аналіз між кількома змінними, побудувати матрицю кореляції та візуалізувати її за допомогою теплової карти.

3. Регресійний аналіз: Студенти мають провести лінійний регресійний аналіз для передбачення значення однієї змінної на основі іншої. Візуалізувати отриману регресійну модель на графіку та проаналізувати похибки прогнозування.

4. Тестування гіпотез: З використанням даних, студенти повинні сформулювати та перевірити статистичну гіпотезу (наприклад, щодо середніх значень двох вибірок), застосовуючи t-тест або інший відповідний тест, і проаналізувати результати.

ЛАБОРАТОРНА РОБОТА 1
ВАРІАЦІЙНІ РЯДИ. НЕЗГРУПОВАНІ РОЗПОДІЛИ ЧАСТОТ

Мета роботи: ознайомитися з використанням незгрупованих емпіричних розподілів частот у практиці аналізу даних.

Основні поняття: емпіричні дані, вибірка, емпіричний розподіл, варіаційний ряд, диференціальні й інтегральні розподіли.

Теоретичні відомості та хід виконання роботи

Основним етапом аналізу даних є збір, класифікація та вивчення їх закономірностей. У ході роботи дослідники оперують поняттями **генеральної сукупності**, до якої належать усі досліджувані об'єкти, та **вибірки (вибіркової сукупності)**, до якої належать об'єкти довільно або випадково відібрані з генеральної сукупності. Вибірка повинна бути репрезентативною, тобто максимально повно й коректно відображати ті властивості генеральної сукупності, що вивчаються в ході дослідження (Огірко & Галайко, 2017).

Залежно від того, чи використовується кількісна чи якісна ознака групування, дані можуть бути представлені як якісні (атрибутивні) або кількісні (варіаційні). Варіація визначає можливість ознаки приймати різні числові значення, які називаються **варіантами**. Таким чином, на початковому етапі статистичного дослідження необхідно створити варіаційний ряд.

Варіаційний ряд – це упорядкований список усіх унікальних значень, що містяться у вибірці або наборі даних, з подальшим визначенням частоти кожного з цих значень (тобто скільки разів вони зустрічаються у вибірці). Це дозволяє отримати уявлення про розподіл значень та їх частоту в досліджуваній групі.

Приклад: Нехай у нас є набір даних про вік учасників дослідження: 25, 30, 28, 35, 30, 25, 28, 32, 30,25. Варіаційний ряд у цьому випадку буде: 25, 28, 30, 32, 35.

Незгруповані варіаційні розподіли застосовують до емпіричних даних, властивості яких виміряні за інтервальними або відносними шкалами і набувають дискретних у вузькому діапазоні значень (Руденко, 2012). У незгрупованому розподілі кожне окреме значення змінної відображається поруч з його частотою (або кількістю випадків) у вибірці. Це може бути вигляд довгого списку чисел та їхніх відповідних частот, або представлення у вигляді таблиці, де перший стовпчик/рядок містить значення, а другий – відповідні частоти.

Наприклад, якщо ми маємо набір даних про вік учасників дослідження, то незгрупований розподіл абсолютних частот може виглядати, як наведено у таблиці.

Вік	25	28	30	32	35
Абсолютна частота	3	2	3	1	1
Відносна частота	0,3	0,2	0,3	0,1	0,1

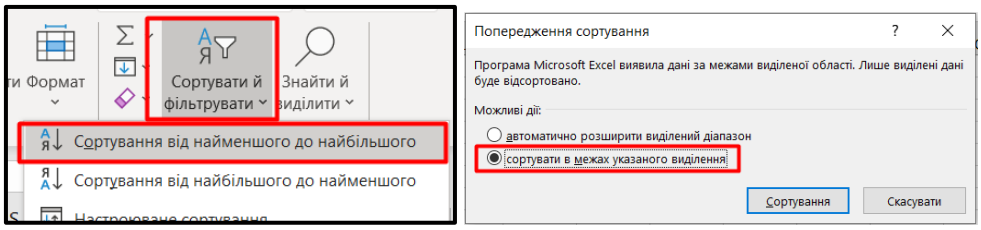
Відношення абсолютної частоти варіанти до об'єму вибірки називають відносною частотою. Залежність між упорядкованим рядом варіант і відповідними їм відносними частотами називається статистичним розподілом відносних частот вибірки.

Розглянемо послідовність виконання розрахунків для незгрупованих розподілів частот засобами MS Excel або Google Таблиць (Руденко, 2012). Інструкції, що відрізняються для кожної з програм, будуть наводитися послідовно.

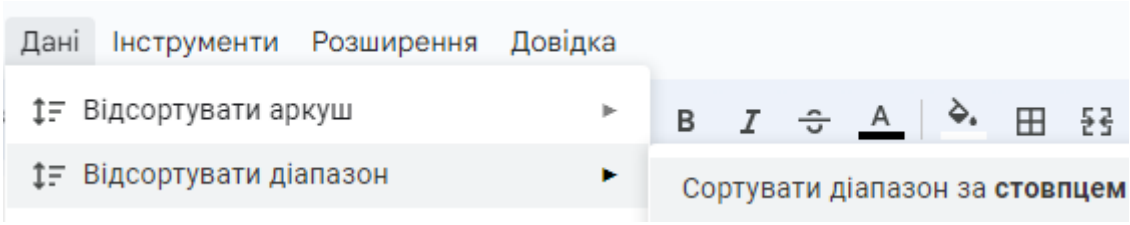
Таблиця містить інформацію про результати тестування 20 студентів. Записуємо значення від 1 до 20 у стовпчик A таблиці, а у стовпчик B записуємо значення, яких набуває змінна x_j.

Кількість правильних відповідей										
j	1	2	3	4	5	6	7	8	9	10
x_j	3	5	4	4	2	6	3	5	4	4
j	11	12	13	14	15	16	17	18	19	29
x_j	4	2	3	5	5	3	4	4	3	5

Впорядковуємо значення за зростанням. У MS Excel для цього можна скопіювати невідсортовані дані в стовпчик C, виділити всі значення мишкою та застосувати операцію **Сортувати й фільтрувати**. Програма може видати попередження. У вікні діалогу слід обрати **Сортувати в межах указаного виділення**.



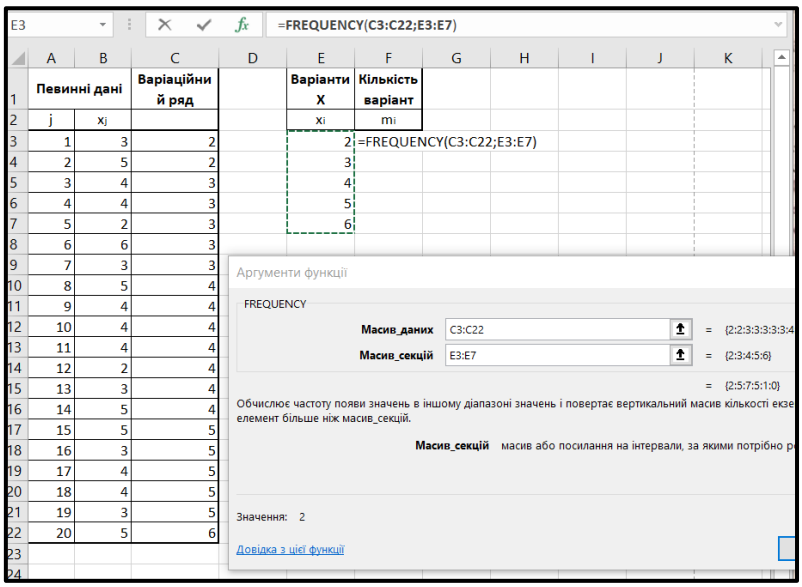
У Google Таблицях виділяєте дані й обираєте **Дані – Відсортувати діапазон.**



Ряди розподілу можуть бути **диференціальними** та **інтегральними** і можуть складатися з абсолютних і відносних частот. Як зазначалося вище, абсолютна частота вказує скільки разів повторюється певна варіанта, а відносна частота – яку частку сукупності складає ця варіанта по відношенню до всієї сукупності.

Диференціальні розподіли описують значення частот окремо (диференційовано) для кожної варіанти. Диференціальні абсолютні частоти – це кількості об'єктів m_i з однаковими значеннями x_i (кількість однакових значень). Диференціальні відносні частоти – це відношення диференціальних абсолютних частот m_i до загальної кількості об'єктів n, тобто, $f_i = m_i/n$.

У комірках E2:E7 записуємо всі можливі варіанти X={2, 3, 4, 5, 6} та визначаємо частоти їх появи m_i.



Для розрахунку частот використовується функція FREQUENCY(). Для її застосування у MS Excel потрібно:

- виділити діапазон F3:F7;
- натиснути F2;
- за допомогою майстра функцій обрати =FREQUENCY();
- задати аргументи функції у вікні діалогу;
- натиснути разом CTRL+SHIFT+ENTER і отримати у комірках F3:F7 значення абсолютних диференціальних частот.

У Google Таблицях вводите в комірці F3 =FREQUENCY(C3:C22;E3:E7).

Для того, щоб розрахувати диференціальні відносні частоти, потрібно поділити кожну з частот m_i на обсяг вибірки $n=20$.

E	F	G
	Диференціальні	
Варіанти X	Кількість варіант (абсолютні)	Відносні
x_i	m_i	f_i
2	=FREQUENCY(C4:C23;E4:E8)	=F4/$F$9
3	=FREQUENCY(C4:C23;E4:E8)	=F5/$F$9
4	=FREQUENCY(C4:C23;E4:E8)	=F6/$F$9
5	=FREQUENCY(C4:C23;E4:E8)	=F7/$F$9
6	=FREQUENCY(C4:C23;E4:E8)	=F8/$F$9
Суми	=SUM(F4:F8)	=F9/$F$9

E	F	G
	Диференціальні	
Варіанти X	Кількість варіант (абсолютні)	Відносні
x_i	m_i	f_i
2	2	0,10
3	5	0,25
4	7	0,35
5	5	0,25
6	1	0,05
Суми	20	1

Інтегральні розподіли («накопичувальні» або «кумулятивні») формуються як доданки попередніх диференціальних частот.

Інтегральні абсолютні частоти $m_j^* = \sum_{i=1}^{j} m_i$ – це накопичена сума диференціальних частот від першої до *j*-ї варіанти. Інтегральні відносні частоти $F_j = \sum_{i=1}^{j} f_i$ – це накопичена суча диференціальних відносних частот від першої до *j*-ї варіанти.

Приклад розрахунків інтегральних абсолютних і відносних частот наведено на рисунку нижче.

E	F	G	H	I
	Диференціальні		Інтегральні	
Варіанти X	Кількість варіант (абсолютні)	Відносні	Абсолютні	Відносні
xi	mi			
2	=FREQUENCY(C4:C23;E4:E8)	=F4/$F$9	=F4	=G4
3	=FREQUENCY(C4:C23;E4:E8)	=F5/$F$9	=H4+F5	=I4+G5
4	=FREQUENCY(C4:C23;E4:E8)	=F6/$F$9	=H5+F6	=I5+G6
5	=FREQUENCY(C4:C23;E4:E8)	=F7/$F$9	=H6+F7	=I6+G7
6	=FREQUENCY(C4:C23;E4:E8)	=F8/$F$9	=H7+F8	=I7+G8
Суми:	=SUM(F4:F8)	=SUM(G4:G8)		

Для перевірки розрахунків потрібно знайти суми частот. Сума диференціальних відносних частот дорівнює одиниці, а сума диференціальних абсолютних частот дорівнює обсягу вибірки.

E	F	G	H	I
	Диференціальні		mi	
Варіанти X	Кількість варіант (абсолютні)	Відносні	Абсолютні	Відносні
x_i	m_i	f_i	m^*_i	F_i
2	2	0,10	2	0,1
3	5	0,25	7	0,35
4	7	0,35	14	0,7
5	5	0,25	19	0,95
6	1	0,05	20	1
Суми	20	1		

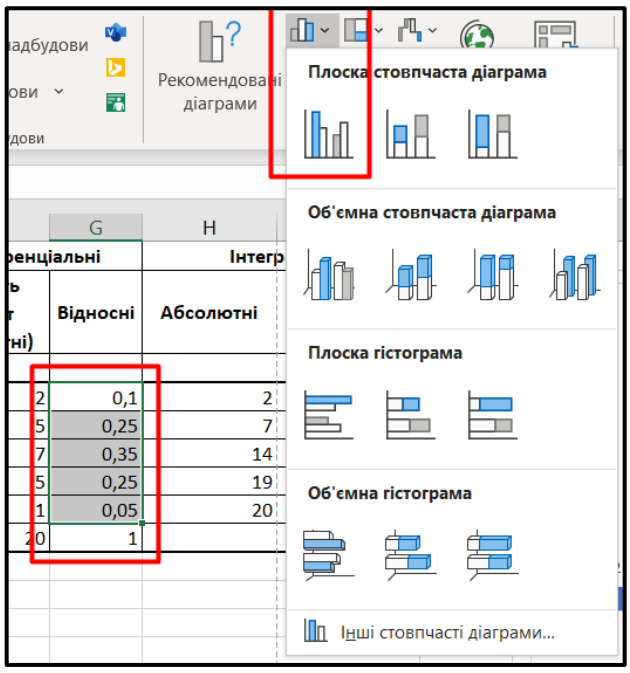
Статистичні розподіли можуть бути представлені у вигляді *аналітичної емпіричної функції розподілу.* Така функція визначає для кожного значення x відносну частоту події.

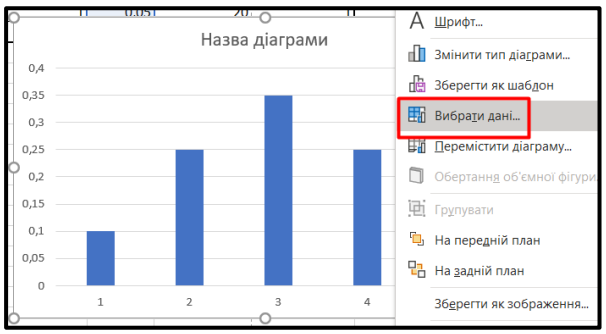
Для нашого прикладу функції диференціального та інтегрального розподілу відносних частот описуються виразами, наведеними нижче.

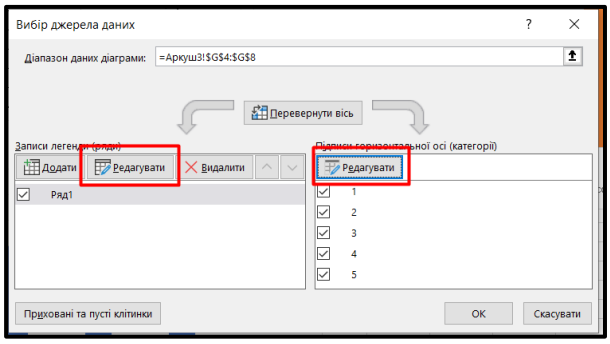
$$f_i(x_i) = \begin{cases} 0{,}00 & x < 2 \\ 0{,}10 & 2 \le x < 3 \\ 0{,}25 & 3 \le x < 4 \\ 0{,}35 & 4 \le x < 5 \\ 0{,}25 & 5 \le x < 6 \\ 0{,}05 & 6 \le x \\ \square & \end{cases} \qquad F_i(x_i) = \begin{cases} 0{,}00 & x_1 < 2 \\ 0{,}10 & 2 \le x < 3 \\ 0{,}35 & 3 \le x < 4 \\ 0{,}70 & 4 \le x < 5 \\ 0{,}95 & 5 \le x < 6 \\ 1 & 6 \le x \\ \square & \end{cases}$$

Графік $F_i(x_i)$ називається функцією розподілу, а $f_i(x_i)$ – функцією густини розподілу. Густина розподілу – це кількість випадків, що припадають на певне дискретне значення ознаки (або інтервал варіювання ознаки).

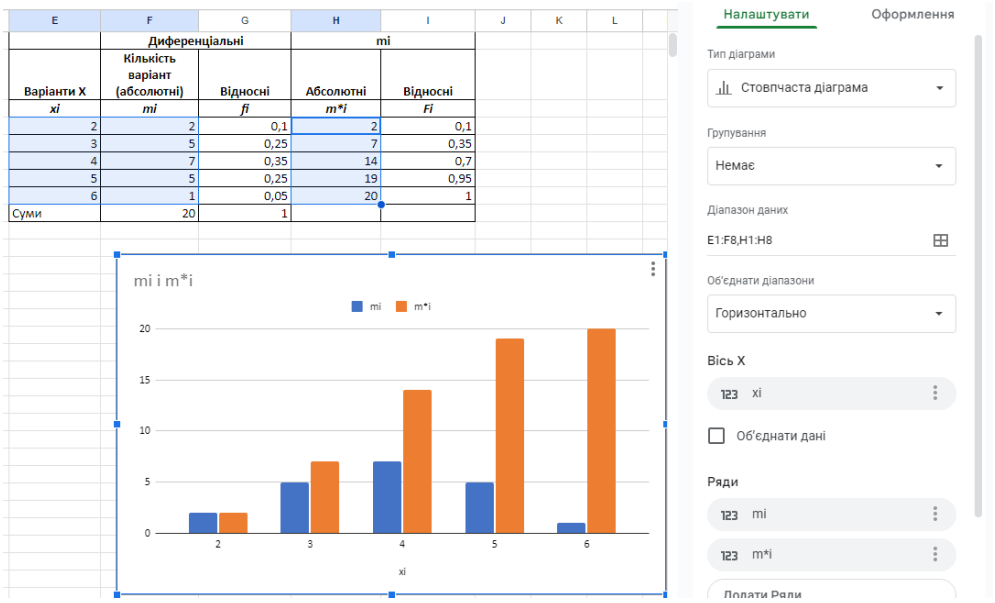
Побудуємо гістограми для диференціального та інтегрального розподілів частот. У MS Excel для цього виділяєте стовпчик даних і (як показано на рисунку) й обираєте тип діаграми. Отриману діаграму редагуєте, за допомогою опції **Вибрати дані**. У випадку незгрупованих розподілів даних, по осі абсцис відкладаються значення варіант, а по осі ординат – абсолютні чи відносні частоти.



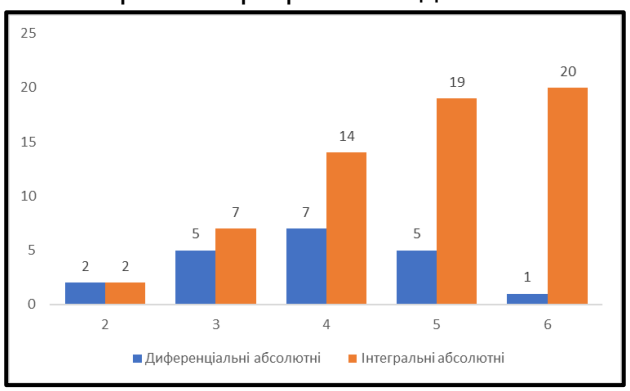


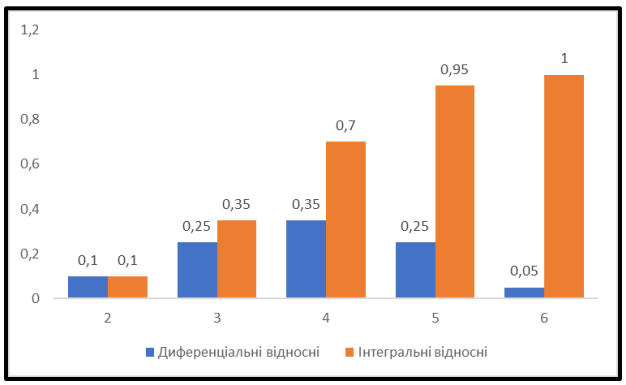


У Google Таблиці, затиснувши клавішу Ctrl, виділяєте мишкою дані у стовпчиках E, F і H. Далі обираєте **Вставити – Діаграма**. За замовчуванням буде додано Стовпчасту діаграму. Тип діаграми й оформлення ви можете змінити за допомогою панелі, що відображена зліва.



Отримані графіки наведено нижче.





Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи.

Для заданої вибірки із генеральної сукупності потрібно здійснити розрахунок диференціальних та інтегральних абсолютних і відносних частот. Побудувати відповідні гістограми незгрупованих розподілів.

Варіант	Завдання
1	126 128 120 124 120 128 126 126 120 124 122 126 122 122 122 124 124 119 119 120 124 126 119 126 124 128 122 125 119 120
2	34 33 36 34 34 32 35 34 34 33 34 34 35 33 34 35 32 32 36 34 34 33 34 32 34 33 35 36 35 32 33 32 36
3	114 114 115 114 114 115 118 115 117 118 115 116 116 116 118 118 117 116 116 116 115 118 117 119 119 115 116 114 116 114
4	383 388 386 386 388 383 386 388 384 385 386 388 385 384 385 386 388 385 386 388 383 386 384 385 385 386 385 388 384
5	159 156 154 158 154 159 157 154 155 156 157 157 158 157 156 159 154 154 159 157 157 154 154 155 156
6	65 68 69 68 67 70 70 64 67 69 71 71 70 69 67 68 66 66 67 65 70 71 71 65 66 68 67 65 66 70 65 70 68 69 67 66 66 65 68 69 66 67
7	146 147 150 148 150 146 148 145 148 150 149 148 149 147 145 150 147 146 150 150 148 147 150 145 146 150 148 150 146 148 150
8	93 95 90 91 94 95 91 92 90 90 91 92 93 93 92 95 91 91 92 91 90 91 94 90 90 90 90 95 94 93 93 92 94 95 92 92 90 91 95
9	20 20 22 20 21 21 25 22 24 21 26 20 22 24 23 22 23 26 24 24 22 23 21 21 20 21 21 20 22 20 22 20 21 24 24 20 20 21 23 22 23 24 24
10	43 44 45 40 41 44 43 43 42 44 45 42 42 40 41 45 44 45 41 42 40 40 41 42 43 43 42 45 41 41 42 41 40 41 44 40 40 40 40 45

За результатами виконання завдання сформувати звіт зі скріншотами обрахунків й отриманих результатів та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА2
ЗГРУПОВАНІ РОЗПОДІЛИ ЧАСТОТ

Мета: ознайомитися з використанням у практиці аналізу даних згрупованих розподілів частот.

Основні поняття: вибірка, варіаційний ряд, інтервальні вимірювання, згрупований розподіл частот.

Теоретичні відомості та хід виконання роботи

Розподіли згрупованих частот використовуються тоді, коли набір даних дуже великий або коли ми хочемо спростити аналіз шляхом групування значень у відповідні інтервали чи категорії. Це особливо корисно, коли маємо справу з неперервними змінними такими як вік, зріст чи вага, де може бути велика кількість унікальних значень.

Також, коли набір даних великий, групування значень дозволяє зменшити кількість окремих категорій, спрощуючи подальший аналіз. Розподіли згрупованих частот дозволяють створити гістограми або полігони частот, які наглядно відображають розподіл даних і дозволяють виявити будь-які відмінності чи тенденції.

Розглянемо послідовність виконання розрахунків для згрупованих розподілів частот засобами електронних таблиць. Потрібно розрахувати розподіли коефіцієнта інтелекту IQ вибірки обсягом у 80 осіб за емпіричними даними у балах (див. таблицю) (Руденко, 2012).

A	B	C	D	E	F	G	H
Результати тестування (80 осіб)							
120	104	102	96	121	97	106	93
83	115	109	119	96	114	91	92
95	112	104	116	85	106	89	102
111	85	113	97	115	105	90	94
92	95	118	104	94	97	109	99
117	97	80	99	86	96	112	102
93	124	98	106	137	93	100	113
120	112	89	78	83	92	72	97
79	80	80	83	87	93	84	87
103	100	107	90	88	105	93	105
Мінімальне		72		**Максимальне**		137	
K=	8	**λ=**	8,125	**Частоти**			
Діапазони значень				диференціальні		інтегральні	
				абсолютна	відносна	абсолютна	відносна
i	Поч.	<IQ≤	Кін.	***m i***	***fi***	***m* i***	***F i***
1	70	<IQ≤	80	6	7,50%	6	7,50%
2	80	<IQ≤	90	14	17,50%	20	25,00%
3	90	<IQ≤	100	26	32,50%	46	57,50%
4	100	<IQ≤	110	16	20,00%	62	77,50%
5	110	<IQ≤	120	15	18,75%	77	96,25%
6	120	<IQ≤	130	2	2,50%	79	98,75%
7	130	<IQ≤	140	1	1,25%	80	100,00%
Суми:				80	1		

Розрахунок здійснюється за наступною послідовністю:

1. Знайти мінімальне й максимальне значення IQ у комірках C12 і G12 за допомогою функцій =MIN(A2:H11) і MAX(A2:H11), отримати відповідно 72 і 137.
2. Розрахувати кількість класів k за формулою Стерджеса $k = 1 + 3{,}32 \cdot \lg n$, де n – обсяг вибірки. Для цього внести у комірку B13 вираз:

=ROUNDUP(1+3,32*LOG10(COUNT(A2:H11));1) і отримати k=8.

3. Розрахувати розмір інтервалу класів $\lambda = \frac{IQ_{max} - IQ_{min}}{k}$ у комірці D13 за допомогою виразу =(G12-C12)/B13. Отримане значення дорівнює 8,125, але з практичної точки зору доцільно взяти розмір інтервалу рівним 10.

4. Встановити границі класів і підрахувати кількість варіант у кожному з них. При підрахунку числа варіант значення, що знаходиться на границі класів, слід відносити завжди до одного й того ж класу, а саме там, де це число трапилося вперше. Відтак воно стає нижньою границею класу. Тому, потрібно розрахувати в комірках A17:D23 значення початкової і кінцевої границь діапазонів значень (кратно 10 балам і так, щоб мінімальне значення 72 входило у перший, а максимальне 137 – в останній інтервал).

5. Для визначення кількості варіант використовується функція =FREQUENCY().

6. Задати аргументи функції, як показано на рисунку.

7. У комірках E17:E23 відобразиться значення абсолютних диференціальних частот.

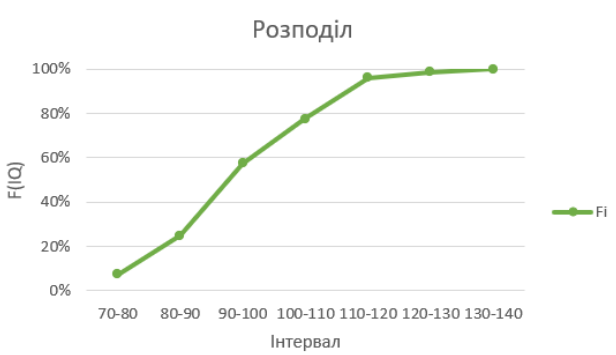
	диференціальні	
	абсолютна	відносна
Кін.		
80	6	
90	14	
100	26	
110	16	
120	15	
130	2	
140	1	

8. Для розрахунку диференціальних відносних, інтегральних абсолютних і відносних частот ввести у комірки F16:H23 відповідні формули.

	A	B	C	D	E	F	G	H
13	K=	=ROUN	λ=	=(G12-(	Частоти			
14	Діапазони значень				диференціальні		інтегральні	
15					абсолютна	відносна	абсолютна	відносна
16	i	Поч.	<IQ≤	Кін.	m	f	m*	f*
17	1	70	<IQ≤	80	=FREQUENCY(A2:H11;D17:D23)	=E17/E$24	=E17	=F17
18	2	80	<IQ≤	90	=FREQUENCY(A2:H11;D17:D23)	=E18/E$24	=E18+G17	=F18+H17
19	3	90	<IQ≤	100	=FREQUENCY(A2:H11;D17:D23)	=E19/E$24	=E19+G18	=F19+H18
20	4	100	<IQ≤	110	=FREQUENCY(A2:H11;D17:D23)	=E20/E$24	=E20+G19	=F20+H19
21	5	110	<IQ≤	120	=FREQUENCY(A2:H11;D17:D23)	=E21/E$24	=E21+G20	=F21+H20
22	6	120	<IQ≤	130	=FREQUENCY(A2:H11;D17:D23)	=E22/E$24	=E22+G21	=F22+H21
23	7	130	<IQ≤	140	=FREQUENCY(A2:H11;D17:D23)	=E23/E$24	=E23+G22	=F23+H22
24	Суми:				=SUM(E17:E23)	=E24/E$24		

9. Отримати результати розрахунку згрупованих частот IQ і побудувати графіки розподілу. Попередньо потрібно відформатувати комірки, обравши відсотковий формат відображення для відносних частот.

E	F	G	H
Частоти			
диференціальні		інтегральні	
абсолютна	відносна	абсолютна	відносна
m	f	m*	f*
6	7,50%	6	7,50%
14	17,50%	20	25,00%
26	32,50%	46	57,50%
16	20,00%	62	77,50%
15	18,75%	77	96,25%
2	2,50%	79	98,75%
1	1,25%	80	100,00%
80	1		



Диференціальний відносний розподіл f(IQ) називається густиною розподілу. Він відображає загальну картину розподілу як усіх категорій разом, так і кожної категорії окремо. Графік розподілу унімодальний і асиметричний, густина концентрується навколо середніх значень. Інтегральний відносний розподіл F(IQ) ілюструє сумарні показники частот для різних діапазонів IQ.

Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи.

Для заданої вибірки із генеральної сукупності здійснити розрахунок диференціальних та інтегральний частот. Побудувати відповідні графіки.

Варіант	Завдання
1	28 38 98 95 84 67 33 103 62 61 97 124 35 67 104 115 71 106 35 76 108 61 118 57 79 94 46 28 112 92 105 116 70 84 111 103 27 76 28 33 72 67 87 41 84 35 91 43 55 113 93 106 53 60 89 39 107 62 99 96 71 53 46 53 103 34 123 41 84 45 35 89 102 94 95 113 37 27 61 75
2	68 62 42 71 89 46 36 24 36 11 78 40 80 60 55 83 90 67 22 68 44 25 88 12 18 63 53 13 53 86 84 69 83 42 56 77 22 22 51 79 12 58 13 84 32 71 88 22 46 40 27 78 81 42 85 71 58 13 72 64 71 19 32 35 75 16 53 52 50 37 25 62 14 36 22 39 78 11 25 21
3	36 27 69 22 71 39 101 23 30 108 106 77 102 99 38 53 24 74 22 69 86 15 92 16 51 37 46 83 83 21 36 101 102 58 88 70 70 18 32 58 53 21 22 20 35 93 79 82 12 19 61 16 81 50 56 105 90 92 17 95 42 45 56 100 30 73 28 104 48 20 85 14 74 93 38 62 95 56 94 59
4	44 108 89 62 77 60 53 77 86 71 62 97 45 63 71 92 39 54 43 70 77 66 72 89 93 47 46 77 92 83 36 66 75 72 49 98 44 50 108 102 80 106 89 89 42 50 103 72 37 108 67 71 83 55 94 45 43 87 74 60 62 107 78 65 63 72 74 71 54 84 76 78 78 37 37 86 78 62 99 53
5	59 25 53 79 64 64 58 39 30 69 76 58 40 23 74 71 84 59 85 38 29 49 79 80 69 44 35 28 85 27 75 68 31 44 50 19 77 22 85 85 78 37 66 22 61 54 65 68 79 23 43 38 40 42 40 36 51 38 69 31 56 30 64 49 50 24 25 35 58 85 72 37 62 77 31 54 29 35 40 74

6	116 56 59 119 86 60 27 25 87 42 75 102 95 41 41 41 98 38 106 67 78 77 41 99 101 50 115 83 116 81 32 103 77 47 110 107 99 87 74 43 120 68 90 38 101 65 51 83 31 112 31 100 90 26 101 109 60 88 120 57 109 55 108 37 48 100 58 53 82 114 29 101 111 85 45 55 25 120 77 58
7	47 36 41 42 30 37 54 120 27 45 30 56 92 115 117 50 42 101 96 108 62 38 43 113 32 82 83 96 39 51 107 44 120 82 25 31 40 106 105 50 80 112 109 72 57 88 55 48 39 26 34 28 99 70 115 91 98 55 114 59 80 71 88 108 114 87 63 101 57 59 53 117 37 73 107 29 50 65 41 68
8	76 92 96 53 76 107 90 77 92 105 83 86 73 105 87 70 94 95 89 82 94 99 100 111 66 78 101 68 64 93 54 65 108 96 64 56 118 43 113 113 102 101 53 104 82 106 61 55 72 90 51 104 118 53 51 115 96 82 95 72 81 94 85 95 120 88 57 50 76 120 80 113 75 118 117 106 55 113 55 46
9	110 80 70 115 52 114 116 71 40 88 49 73 32 36 117 67 92 45 76 56 99 48 90 118 45 88 54 120 62 113 72 104 33 61 63 90 65 56 40 36 98 64 115 94 57 39 59 96 75 38 105 37 47 100 63 104 65 39 46 116 116 60 116 72 86 68 42 76 77 48 91 77 47 94 118 66 40 112 51 116
10	41 112 70 98 109 82 94 96 98 38 51 73 41 46 75 35 98 74 36 102 80 75 112 108 106 113 83 97 53 67 66 35 75 106 62 68 60 106 68 69 60 83 75 96 104 37 44 40 91 100 113 42 65 64 76 54 44 101 62 87 89 42 104 63 108 108 102 46 42 106 89 94 88 112 94 43 42 44 39 42

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА 3
АТРИБУТИВНІ ТА РАНЖИРУВАНІ РОЗПОДІЛИ

Мета: ознайомитися з можливостями використання.MS Excel для побудови й аналізу атрибутивних і ранжируваних розподілів

Основні поняття: атрибутивні розподіли, ранжирувані розподіли.

Теоретичні відомості та хід виконання роботи

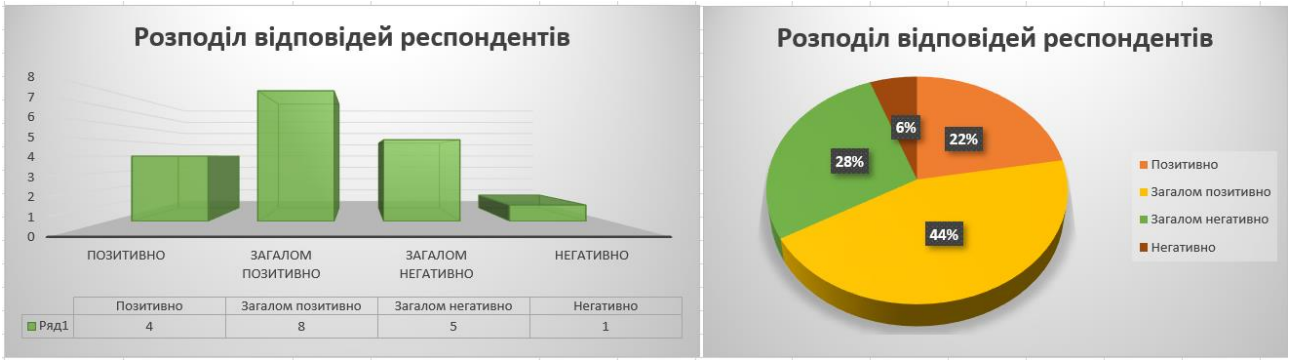
Атрибутивні розподіли використовуються у випадку номінальних (категоріальних) типів вимірювань властивостей досліджуваних об'єктів.

Розглянемо приклад побудови атрибутивного розподілу за результатами опитування глядачів про їх враження від фільму. Респонденти могли обрати один з чотирьох варіантів відповіді – «негативно», «загалом негативно», «загалом позитивно», «позитивно». Усього в дослідженні брали участь 16 осіб. Отримані результати збережені в таблиці (стовпчики A:D). Відповідно до відповідей респондентів, кожній особі надано відповідний атрибут x_i, наприклад, «позитивно» - 1, «загалом позитивно» - 2 і т.д. (стовпчики C:D і E:F).

	A	B	C	D	E	F	G	H
1	№ з/п	Респонденти	Результати опитування		Типи ВНД		Частоти	
2	j		xj		xi		mi	mi/n
3	1	Андрій	Позитивно	1	Позитивно	1	4	22,22%
4	2	Тарас	Загалом позитивно	2	Загалом позитивно	2	8	44,44%
5	3	Анна	Загалом негативно	3	Загалом негативно	3	5	27,78%
6	4	Олександр	Позитивно	1	Негативно	4	1	5,56%
7	5	Юлія	Загалом позитивно	2	Загалом:		18	100%
8	6	Марина	Загалом негативно	3				
9	7	Оксана	Негативно	4				
10	8	Вадим	Позитивно	1				
11	9	Віталій	Загалом позитивно	2				
12	10	Ростислав	Загалом позитивно	2				
13	11	Яна	Загалом негативно	3				
14	12	Катерина	Загалом позитивно	2				
15	13	Софія	Загалом позитивно	2				
16	14	Анастасія	Загалом негативно	3				
17	15	Влада	Загалом позитивно	2				
18	16	Дмитро	Загалом негативно	3				
19	17	Ігор	Загалом позитивно	2				
20	18	Вікторія	Позитивно	1				

Для розрахунку абсолютних частот mi у комірку G3 потрібно записати вираз =COUNTIF($D$3:$D$18;F3). Аналогічні вирази запишіть у комірки G4:G6. Для розрахунку загальної кількості об'єктів n у комірці G7, використайте функцію сумування. Для розрахунку відносних частот $\frac{m_i}{n}$ у комірку H3 запишіть функцію =G3/$G$7, аналогічні вирази запишіть у комірки H4:H6.

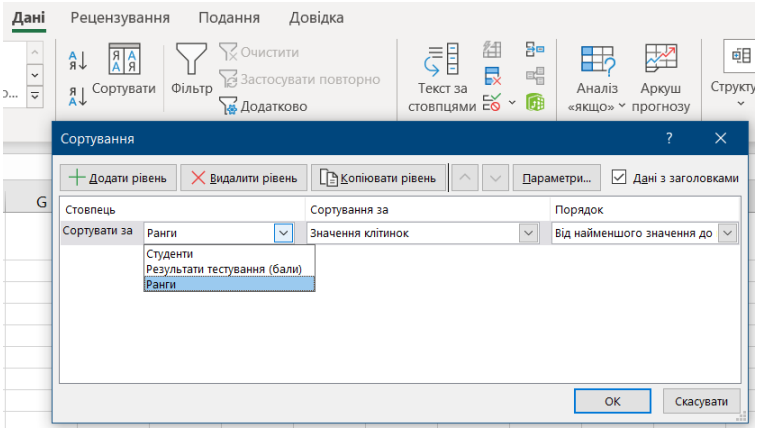
Для ілюстрації атрибутивних розподілів використовують два найпоширеніші типи графіків: гістограму та колову діаграму. Атрибутивні розподіли дозволяють оцінити властивості в абсолютних і відносних значеннях, наприклад, співвідношення різних характеристик. Аналіз побудованих графіків дозволяє стверджувати, що більшості респондентів фільм сподобався.



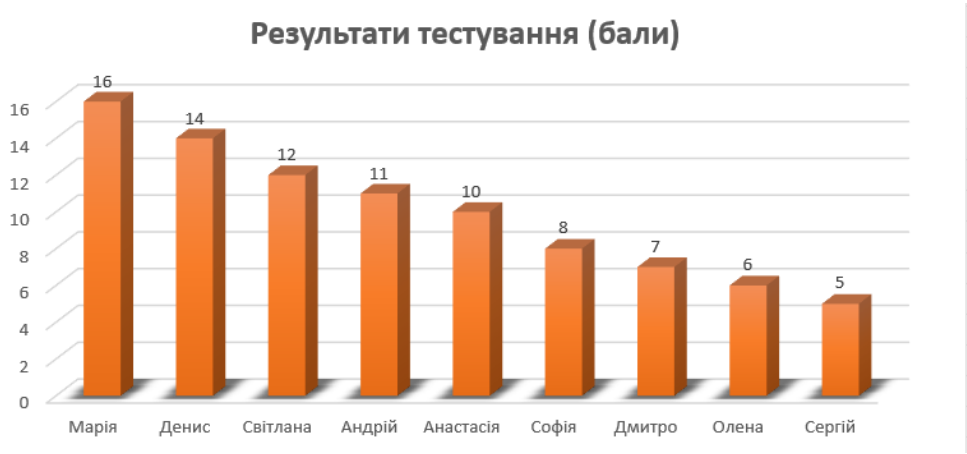
Ранжирувані розподіли використовують у разі порядкових (рангових) типів вимірювань, наприклад, визначення рейтингу успішності певної діяльності. Розглянемо приклад ранжування студентів за результатами екзаменаційного тестування (стовпчики A:B) (Руденко, 2012). Спочатку визначимо ранг значення кожного з результатів серед даних вибірки у діапазоні $B$2:$B$10.

A	B	C	C
Студенти	Результати тестування (бали)	Ранги	Ранги
Олена	6	=RANK.AVG(B2;$B$2:$B$10)	8
Андрій	11	=RANK.AVG(B3;$B$2:$B$10)	4
Марія	16	=RANK.AVG(B4;$B$2:$B$10)	1
Анастасія	10	=RANK.AVG(B5;$B$2:$B$10)	5
Сергій	5	=RANK.AVG(B6;$B$2:$B$10)	9
Дмитро	7	=RANK.AVG(B7;$B$2:$B$10)	7
Денис	14	=RANK.AVG(B8;$B$2:$B$10)	2
Софія	8	=RANK.AVG(B9;$B$2:$B$10)	6
Світлана	12	=RANK.AVG(B10;$B$2:$B$10)	3
Сума рангів:		=SUM(C2:C10)	45

Далі потрібно упорядкувати дані діапазону A2:C10 за рангом за допомогою команди Дані – Сортування і відобразити отриманий рейтинг графічно.



Ранжирувані розподіли дозволяють візуалізувати результати дослідження певної властивості серед об’єктів щодо їх зростання чи спадання. Вони характеризують усю сукупність та кожну її одиницю окремо.



Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи.

Завдання 1. Атрибутивний розподіл. Відповідно до визначеної тематик, створіть випадковим чином таблицю відповідей 20 респондентів. Розрахуйте відносні й абсолютні частоти, побудуйте гістограму та колову діаграму.

№ з/п	Завдання
1	Пори року (зима, весна, літо, осінь)
2	Кольори (червоний, зелений, синій, жовтий)
3	Планети (Юпітер, Сатурн, Уран, Нептун)
4	Спеціальності (СОІ, АКІТ, ПМ, ПФ)
5	Країни (Франція, Італія, Іспанія, Португалія)
6	Фрукти (вишня, яблука, сливи, абрикоси)
7	Місяці (січень, березень, червень, вересень)
8	Планети (Меркурій, Венера, Земля, Марс)
9	Кольори (чорний, білий, зелений, червоний)
10	Міста (Рим, Париж, Лондон, Мадрид)

Завдання 2. Ранжируваний розподіл. У ході дослідження збиралися відгуки експертів стосовно характеристик деякого програмного сервісу. Розрахуйте середні оцінки для кожної з характеристик, визначте ранг кожної з них. Побудуйте діаграму для впорядкованих за рангом середніх значень оцінки (за характеристиками).

A	B	C	D	E	F	G	H
	Експерти					Середня	
Характеристики	1	2	3	4	5	оцінка	Ранг
Зручність	8	5	6	4	6		
Змістовність	1	6	2	8	7		
Функціональність	7	1	3	1	1		
Структурованість	6	4	4	2	2		
Актуальність	2	3	1	3	3		
Швидкодія	3	2	5	5	4		
Дизайн	5	8	8	6	8		
Кольорова гама	4	7	7	7	5		

ЛАБОРАТОРНА РОБОТА 4
РОЗРАХУНОК СТАТИСТИЧНИХ ПАРАМЕТРІВ ВИБІРОК

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для розрахунку мір центральної тенденції та мір мінливості вибірок.

Основні поняття: вибірка, основні параметри описової статистики, середня арифметична величина, мода, медіана, дисперсія, стандартне (середнє квадратичне) відхилення.

Теоретичні відомості та хід виконання роботи

Для отримання детальної інформації про експериментальні дані обчислюють значення ***статистичних параметрів*** вибірки – *середнє арифметичне, моду, медіану, дисперсію, стандартне відхилення* (Огірко & Галайко, 2017).

Центральна тенденція – це оцінка центру розподілу змінної. Існує три оцінки мір центральної тенденції (МЦТ): середнє арифметичне, медіана та мода.

Середнє арифметичне – це узагальнююча величина, що характеризує рівень варіюючої ознаки в якісно однорідній сукупності.

Для обчислення значення середньої арифметичної величини вибірки, дані якої не впорядковані використовують формулу:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1+x_2+\cdots+x_n}{n} \quad (4.1)$$

де x_i – різні варіанти вибірки від x_1 до x_n; n – обсяг вибірки.

Якщо дані вибірки є впорядкованими і нам відомі частоти n_i, то для обчислення середнього вибірки використовується формула:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n} = \frac{x_1 \cdot n_1+x_2 \cdot n_2+\cdots+x_k \cdot n_k}{n} \quad (4.2)$$

де x_i – різні варіанти вибірки, впорядковані за зростанням від x_1 до x_n; k – кількість різних значень варіант вибірки; n_i – частоти кожного із варіант вибірки (числа, що вказують на кількість повторів кожного із представників x_i), n – обсяг вибірки.

Медіана – це таке значення x_i, що знаходиться в середині варіаційного ряду. Фактично, медіана – це значення, яке ділить вибірку навпіл. Тобто 50% значень є меншими від медіани і 50% - більшими.

Якщо число, що позначає кількість випробуваних є парним, то числове значення *медіани* (M_e) обчислюється як половина суми двох значень x_i вибірки з порядковими номерами $n/2$ і $(n+2)/2$:

$$M_e = \frac{x_{n/2}+x_{n+2/2}}{2} \quad (4.3)$$

Якщо число, що позначає кількість випробуваних, є не парним, то числове значення *медіани* (M_e) обчислюється за формулою:

$$M_e = x_{\frac{n+1}{2}} \quad (4.4)$$

Модою вибірки *(M_o)* називають значення варіанти, що найчастіше зустрічається у вибірці, тобто значення x_i з найбільшою частотою n_i і для дискретних рядів вона визначається як $M_o = x_i$, якщо $n_i = max n_i$.

Міри мінливості включають *дисперсію вибірки, стандартне відхилення, коефіцієнт варіації, асиметрію, ексцес*.

Дисперсія – це величина, за допомогою якої характеризують розсіювання або скупченість навколо центру розсіювання статистичних даних. Це середнє арифметичне квадратів відхилень кожного значення ознаки від середньої величини.

Дисперсія вибірки впорядкованих даних обчислюється за допомогою формули:

$$s_x^2 = D_{\text{В}} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}{n} = \frac{(x_1 - \bar{x})^2 \cdot n_1 + (x_2 - \bar{x})^2 \cdot n_2 + \cdots + (x_k - \bar{x})^2 \cdot n_k}{n} \quad (4.5)$$

Середнє квадратичне (стандартне) відхилення – це величина, за допомогою якої характеризують розсіювання або скупченість навколо центра розсіювання статистичних даних. Ця величина показує на скільки в середньому кожна варіанта вибірки відрізняється від середньої арифметичної величини цієї вибірки.

Стандартне відхилення обчислюється за допомогою формули:

$$s_x = \sqrt{s_x^2} = \sqrt{D_{\text{В}}}. \quad (4.6)$$

Значення цього параметра характеризує мінливість ознаки як в сторону збільшення варіант від значення середньої арифметичної величини, так і в сторону зменшення. Значення стандартного відхилення вимірюється в тих самих величинах, що й ознака, тобто величина, щодо якої досліджують вибірку.

Значення *коефіцієнта варіації* показує, яку частину середнє квадратичне відхилення складає від значення середньої арифметичної величини і характеризує степінь мінливості у відсотках.

$$V = \frac{s_x}{\bar{x}} \cdot 100\%$$

Чим більше величина коефіцієнта варіації, тим більш мінлива, неоднорідна ознака. У залежності від значення величини розрізняють:

- ознаки з низькою неоднорідністю (будемо вважати вибірку однорідною) за умови, що $V = C_V$ = 1% − 15%;
- ознаки з середньою неоднорідністю (будемо вважати вибірку з середнім показником розсіювання) $V = C_V$ = 15,1% − 25%;
- ознаки з високою неоднорідністю, розсіюванням, за умови, що $V = C_V \geq 25\%$.

Розглянемо способи розрахунку МЦТ і ММ з використанням властивостей табличних процесорів MS Excel і Google Таблиць.

Спосіб 1

Для невпорядкованих даних перший спосіб обрахунку передбачає використання вбудованих функцій табличних процесорів. Перелік статистичних параметрів та їх короткий опис наведено в таблиці 4.1.

Таблиця 4.1

Вбудовані статистичні функції MS Excel та Google Таблиць

Міри центральної тенденції (МЦТ)			
1.	Мода	Значення, яке найчастіше трапляється серед даних вибірки.	*MODE.SNGL()*
2.	Медіана	Значення, яке припадає на середину впорядкованої вибірки.	*MEDIAN()*
3.	Середнє	Сума всіх значень вибірки, поділена на їх кількість	*AVERAGE()*

Міри мінливості (ММ)			
1.	Дисперсія	Величина, за допомогою якої характеризують розсіювання або скупченість навколо центра розсіювання статистичних даних	*VAR.S()*
2.	Стандартне відхилення	Величина, що показує наскільки в середньому кожна варіанта вибірки відрізняється від середньої арифметичної величини цієї вибірки	*STDEV.S()*
3.	Асиметрія	Характеризує ступінь несиметричності розподілу, відносно його середнього значення (див. рис. нижче).	*SKEW()*
4.	Ексцес	Характеризує відносну опуклість або зглаженість розподілу вибірки, порівняно з нормальним розподілом (див. рис. нижче).	*KURT()*

Приклад розрахунків статистичних параметрів для невпорядкованої вибірки наведено на рисунку.

	A	B
1	**Емпіричні дані**	
2	**i**	**xi**
3	1	1
4	2	6
5	3	2
6	4	7
7	5	5
8	6	2
9	7	6
10	8	8
11	9	2
12	10	8
13	11	8
14	12	1
15	13	7
16	14	4
17	15	2
18	16	3
19	17	8
20	18	5
21	19	1
22	20	8

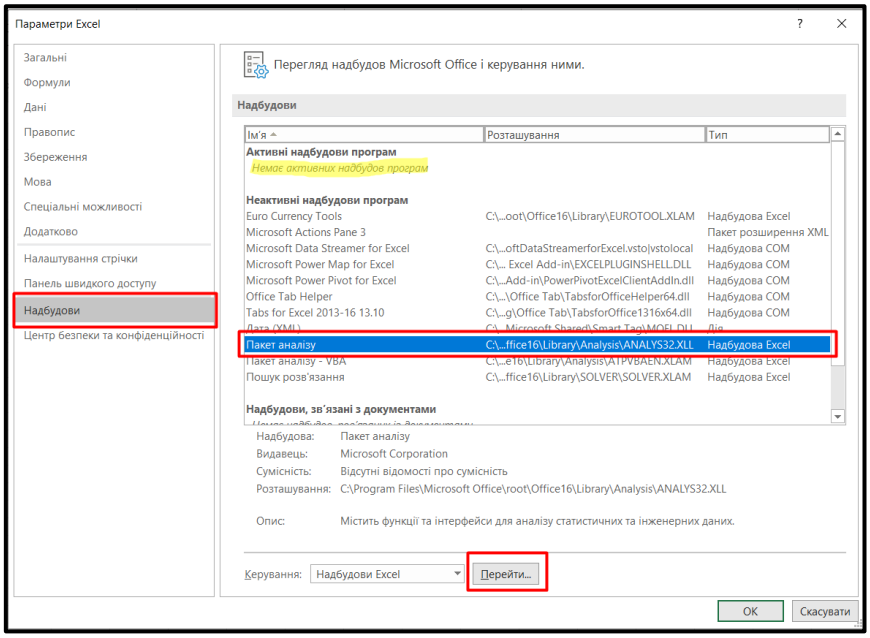
МЦТ	
Мода	8
Медіана	5
Середнє	4,7
ММ	
Дисперсія	7,4842105
Стандартне відхилення	2,7357285
Ексцес	-1,684559
Асиметрія	-0,078488

МЦТ	
Мода	=MODE.SNGL(B3:B22)
Медіана	=MEDIAN(B3:B22)
Середнє	=AVERAGE(B3:B22)
ММ	
Дисперсія	=VAR.S(B3:B22)
Стандартне відхилення	=STDEV.S(B3:B22)
Ексцес	=KURT(B3:B22)
Асиметрія	=SKEW(B3:B22)

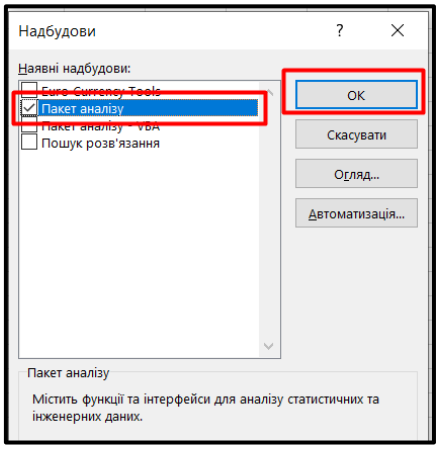
Спосіб 2

Другий спосіб розрахунків можливо реалізувати лише в MS Excel, оскільки він ґрунтується на використанні пакету «Data Analysis» («Аналіз даних») (розділ «Descriptive Statistics» («Описова статистика»)).

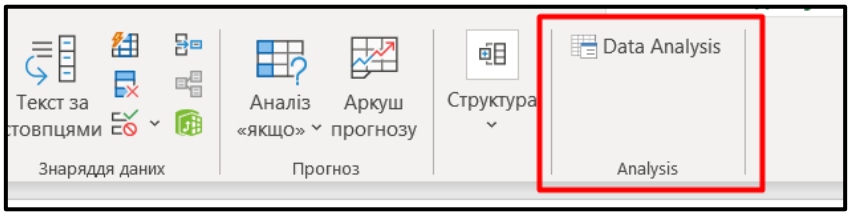
Для активації пакету потрібно обрати у меню **Файл – Параметри – Надбудови** **Пакет аналізу** (на рисунку показано, що жоден з пакетів ще не є активованим). Далі потрібно натиснути Перейти.



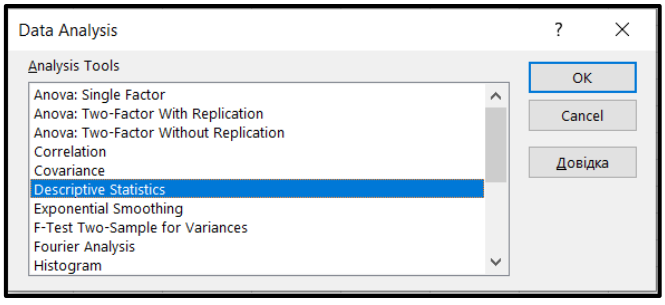
У вікні діалогу обрати **Пакет аналізу** і натиснути OK.



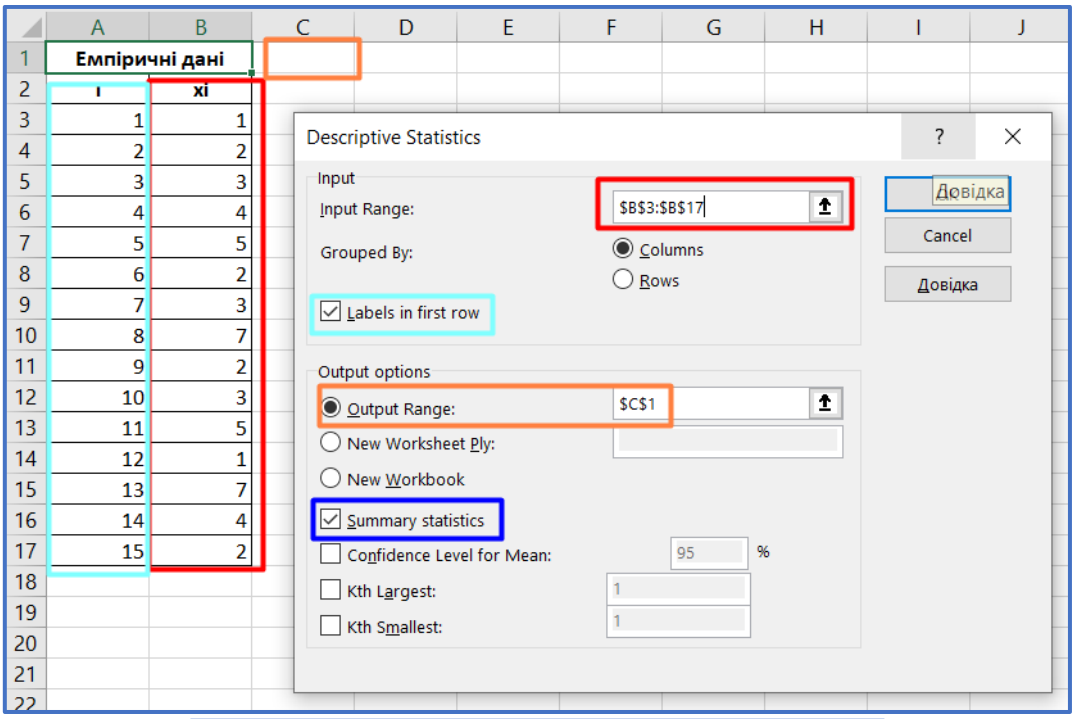
На вкладці Дані відобразиться кнопка Data Analysis.



Для аналізу вибірки потрібно натиснути Data Analysis. Серед можливостей, що пропонуються обрати «Descriptive Statistics» («Описова статистика»).



Приклад даних, налаштування параметрів описової статистики й отримані результати показано на рисунках.



Емпіричні дані		1	
i	xi		
1	1	Mean	3,571428571
2	2	Standard Error	0,499607381
3	3	Median	3
4	4	Mode	2
5	5	Standard Deviation	1,869359648
6	2	Sample Variance	3,494505495
7	3	Kurtosis	-0,301379764
8	7	Skewness	0,735125441
9	2	Range	6
10	3	Minimum	1
11	5	Maximum	7
12	1	Sum	50
13	7	Count	14
14	4		
15	2		

У ситуації, коли нам відомий розподіл частот вибірки, для розрахунку статистичних параметрів потрібно послуговуватися формулами (4.2), (4.5), (4.6).

Розглянемо розрахунок середнього, дисперсії та стандартного відхилення на приладі дискретної вибірки, що задана таблицею.

x_i	-1	2	5	8	10
n_i	5	10	20	5	10

Записуємо розрахункові формули у відповідні комірки. Зазначимо, що моду вибірки можна знайти визначивши максимальне значення частоти. У наведеному випадку це 20, тому мода рівна 5.

	A	B	C	D
1	**xi**	**ni**	**xi*ni**	ni*xi^2
2	-1	5	=A2*B2	=B2*A2^2
3	2	10	=A3*B3	=B3*A3^2
4	5	20	=A4*B4	=B4*A4^2
5	8	5	=A5*B5	=B5*A5^2
6	10	10	=A6*B6	=B6*A6^2
7		**=SUM(B2:B6)**	**=SUM(C2:C6)**	**=SUM(D2:D6)**
8				
9	Середнє:	=C7/B7		
10	Квадрат середнього:	=B9^2		
11	Дисперсія:	=D7/B7-B10		
12	Стандартне відхилення:	=SQRT(B11)		

Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи.

1. Використовуючи вбудовані функції табличних процесорів, розрахувати МЦТ та ММ для наведеної вибірки.

Варіант	Завдання
1	Проводилося дослідження, під час якого фіксувалася середньодобова температура повітря (у Цельсіях): 12,6 12,8 12,0 12,4 12,0 12,8 12,6 12,5 12,0 12,4 12,2 12,6 12,2 12,2 12,2 12,4 12,4 11,8 11,9 12,0 12,4 12,6 11,9 12,6 12,4 12,8 12,2 12,5 11,9 12,0
2	Здійснювалося вимірювання тиску в посудині (у МПА): 3,41 3,32 3,62 3,47 3,42 3,29 3,58 3,42 3,47 3,32 3,42 3,47 3,58 3,32 3,42 3,58 3,25 3,29 3,62 3,42 3,47 3,32 3,41 3,29 3,47 3,58 3,42 3,47 3,44 3,52 3,47 3,55
3	У дослідженні фіксувалися результати вимірювання напруги (у Вольтах): 11,47 12,50 11,48 11,52 11,48 11,48 11,53 11,48 11,52 11,57 11,58 11,59 11,60 11,62 11,63 11,62 11,59 11,67 11,63 11,68 11,62 11,69 11,68 11,67 11,64 11,65 11,65 11,67 12,67 11,56 12,60 12,68 12,56 12,57 11,57 12,65
4	Проводилося дослідження, за якого фіксувалися температурні показники (у Цельсіях). 38,2 37,8 36,6 37,6 38,9 40,0 37,6 37,8 39,0 40,0 37,6 40,0 36,8 38,5 39,0 39,0 36,6 37,8 40,5 37,6 37,8 39,0 36,6 39,0 40,0 40,0 37,6 40,0 37,8 39,0.
5	Проводилося дослідження, за якого визначали швидкість протікання хімічної реакції. Отримані наступні дані (час у секундах): 15,9 17,6 17,4 16,0 18,4 15,9 18,7 18,4 17,4 17,6 19,2 19,4 18,2 17,5 18,4 15,9 17,4 18,4 17,9 20,0 18,7 17,4 18,4 18,4.
6	Проводилося дослідження, під час якого визначали зміну сили струму в установці (в Амперах): 65,0 68,2 69,0 68,3 67,2 70,0 72,2 67,2 69,4 71,2 71,3 67,4 70,2 71,5 70,6 66,3 71,0 68,2 67,6 69,1 71,4 68,5 67,5 70,5 71,5 70,6

7	Під час дослідження вимірювалася ширина виробу (у мм) : 14,5 15,3 15,0 14,8 15,0 15,3 14,8 15,7 14,8 15,0 15,3 15,7 15,3 15,7 14,5 15,0 15,3 14,5 15,7 15,3 14,8 15,7 15,0
8	У результаті статистичних досліджень відгуків експертів, щодо властивостей системи, отримано: 93,5 90,0 91,5 91,5 92,0 90,5 90,0 89,5 91,5 91,0 92,5 91,0 90,5 91,5 90,0 90,0 90,5 90,2 92,3 92,4
9	Проводились змагання з метання ядра. Отримали наступні результати 22 легкоатлетів (м): 17,5 20,2 20,0 21,0 21,5 22,0 21,6 20,8 20,5 21,0 21,5 20,5 22,0 20,8 22,0 20,8 21,6 20,2 20,5 21,0 20,0 21,5.
10	У результаті контрольних вимірювань однотипних деталей, було отримано масив даних: 95,0 93,5 97,5 90,0 100,5 95,0 93,5 101,0 95,0 97,5 93,5 95,0 97,5 90,0 99,5 95,0 97,5 93,5 95,0 100,5 97,5 93,5 95,0 92,5 99,5.

2. Для даних, заданих таблицями, розрахувати середнє значення, дисперсію та стандартне відхилення.

Варіант	**Завдання**					
1	x_i	2	5	7	8	10
	n_i	6	4	5	5	5
2	x_i	-2	0	3	5	7
	n_i	7	13	14	15	11
3	x_i	-2	-1	0	1	2
	n_i	1	3	2	4	2
4	x_i	2	3	5	6	8
	n_i	15	10	25	30	35
5	x_i	2	4	6	8	10
	n_i	2	3	5	1	4
6	x_i	6	8	10	12	14
	n_i	5	1	4	2	3
7	x_i	3	5	7	10	15
	n_i	2	4	7	4	3
8	x_i	-2	-1	0	1	2
	n_i	7	13	14	15	11
9	x_i	2	3	5	6	7
	n_i	6	4	5	5	3
10	x_i	4	8	12	16	20
	n_i	1	3	5	4	4

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 5
ДОВІРЧІ ІНТЕРВАЛИ Й ДОВІРЧА ІМОВІРНІСТЬ

Мета: ознайомитися з поняттям довірчих інтервалів та можливостями використання табличних процесорів для їх розрахунку.

Основні поняття: довірчий інтервал, довірча ймовірність, генеральна сукупність, вибірка.

Теоретичні відомості та хід виконання роботи

Одним із завдань математичної статистики є оцінка числових характеристик генеральної сукупності за вибірковими даними. **Генеральною сукупність** є набором об’єктів, з яких випадковим чином формується вибірка. Наприклад, якщо зі 1000 студентів ми аналізуємо оцінки 100, то об’єм генеральної сукупності N=1000, а об'єм вибірки – n=100. Зазвичай, дослідники мають інформацію лише для вибірок. Зазначимо, що вибірка повинна достатньо добре відтворювати властивості генеральної сукупності, тобто бути представницькою або репрезентативною. Наприклад, якщо в складі досліджуваної сукупності присутніми є 600 осіб жіночої і 400 – чоловічої статі, то репрезентативна вибірка обсягом у 100 студентів повинна зберегти пропорційне (60% і 40%) представництво осіб кожної статі.

Для вибірки можна розрахувати вибіркове середнє, моду, медіану, вибіркову дисперсію та вибіркове середньоквадратичне відхилення. Теоретичну основу оцінювання з використанням вибіркового методу складає закон великих чисел, згідно з яким при необмеженому збільшення об’єму вибірки, випадкові характеристики вибірки наближаються до відповідних параметрів генеральної сукупності.

Якщо об’єм вибірки є незначним, для підвищення точності обробки даних використовуються інтервальні оцінки. Інтервальною називається оцінка, що визначається двома числами – кінцями інтервалів.

Довірчим інтервалом для певного параметру генеральної сукупності називається такий числовий інтервал, у межах якого знаходиться цей параметр. Найчастіше довірчий інтервал обирається симетричним, до досліджуваного параметру. Імовірність, з якою довірчий інтервал охопить істинне значення параметра, називається **довірчою ймовірністю** або **рівнем надійності**. Розмір довірчого інтервалу залежить від обсягу вибірки n (зменшується зі збільшенням n) і від значення довірчої ймовірності (збільшується при її наближенні до одиниці).

Значення довірчої імовірності обирає дослідник, залежно від того, якої точності вимагає дослідження. Події з імовірністю, близькою до 1, вважаються вірогідними (достовірними), а події з імовірністю, близькою до 0, визнаються невірогідними (неможливими). Зазвичай, ці значення довірчої імовірності знаходяться в інтервалі від 0,9 до 0,999.

Поруч із поняттям «довірча імовірність» (як правило, позначається Θ) використовується поняття «рівень значущості» (α або γ): Θ=1-α. Рівень значущості вказує ймовірність помилки оцінювання.

Довірчі інтервали розраховуються з урахуванням певних вимог до генеральної сукупності (вимога нормальності розподілу даних).

Для нормального розподілу модель інтервальної оцінки середнього генеральної сукупності μ має вигляд:

$$\mu \in [\bar{X} - \Delta, \bar{X} + \Delta]$$

де $\Delta = \frac{z_{\alpha/2} \cdot s_x}{\sqrt{n}}$, $\bar{X}$ і s_x – вибіркове середнє і стандартне відхилення, n – обсяг вибірки, $z_{\alpha/2}$ – параметр стандартного нормального розподілу, α – рівень значущості (імовірність того, що відхилення вибіркового від генерального середнього не перевищить Δ за абсолютним значенням.

У психологічних і педагогічних дослідженнях загальноприйнятими вважаються так звані стандартні значення Θ і α (див. табл. нижче). При двосторонній критичній області критичні точки визначають для ймовірності α⁄2, а при 1-сторонній критичній області – для ймовірності α.

Таблиця 5.1

Стандартні значення довірчої ймовірності Θ, рівня значущості α і параметра z

Довірча ймовірність	**Рівень значущості**	**Параметр нормального розподілу**	
Θ	α	z_α	$z_{\alpha/2}$
0,90 (90% вірогідності)	0,10	1,28	1,64
0,95 (95% вірогідності)	0,05	1,64	1,96
0,99 (99% вірогідності)	0,01	2,33	2,58
0,999 (99,9% вірогідності)	0,001	3,09	3,29

5.1. Побудова довірчого інтервалу при відомомих значеннях середнього арифметичного та стандартного відхилення

Приклад. Вибірка обсягом 80 осіб має середнє арифметичне $\bar{X} = 100$ і стандартне відхилення $s_x = 5{,}6$. Необхідно оцінити довірчий інтервал для середнього генеральної сукупності μ на рівні значущості 0,05 (для нормального розподілу $z_{\alpha/2}$).

Послідовність рішення.

1. Визначити параметр стандартного нормального розподілу $z_{\alpha/2}$ для рівня значущості α за допомогою вбудованої функції =NORM.S.INV(1-0,05/2), яка виводить значення 1,96. Отримане значення відповідає табличному.

Функція NORM.S.INV(імовірність) повертає обернене значення стандартного нормального інтегрального розподілу. Цей розподіл має середнє, що дорівнює нулю, і стандартне відхилення, що дорівнює одиниці. Аргументом функції NORM.S.INV є значення ймовірності при $\alpha/2$ $(\Theta = 1 - \alpha/2)$.

2. Довірчий інтервал середнього генеральної сукупності визначаємо двома способами:

2.1. За формулою, наведеною вище.

$$\Delta = \frac{z_{\alpha/2} \cdot s_x}{\sqrt{n}} = \frac{1{,}96 \cdot 5{,}6}{\sqrt{80}} = 1{,}23.$$

2.2. Використовуючи функцію =CONFIDENCE.NORM(0,05;5,6;80)=1,23 з синтаксисом CONFIDENCE.NORM(альфа;станд_відхилення;розмір).

3. Визначаємо межі довірчого інтервалу $[\bar{X} - \Delta, \bar{X} + \Delta]$.

Таким чином, на рівні значущості 0,05 середнє генеральної сукупності належить діапазонові 100,0+/-1,23. Інакше кажучи, з довірчою імовірністю 95% середнє покривається діапазоном значень від 98,77 до 101, 23.

	A	B	C	D	E
1	n	80		Параметр z(α/2)	=NORM.S.INV(1-B4/2)
2	Середнє	100		Спосіб 1	=E1*B3/SQRT(B1)
3	Стандартне відхилення	5,6		Спосіб 2	=CONFIDENCE.NORM(B4;B3;B1)
4	Рівень значущості	0,05		Межа 1	=B2-E3
5	Довірча ймовірність	0,95		Межа 2	=B2+E3

5.2. Побудова довірчого інтервалу при невідомому значенні дисперсії генеральної сукупності із заданою надійністю.

Для вибірок незначних за розміром для оцінювання середнього для генеральної сукупності при невідомому значенні дисперсії генеральної сукупності нгеможливо скористатися нормальним законом розподілу. Тому, для довірчого інтервалу використовується випадкова величина, що має розподіл Стьюдента з $k = n - 1$ ступенями свободи:

$$t = \frac{\bar{X} - \alpha}{\frac{S}{\sqrt{n}}}$$

Приклад. Випадково вибрана партія з двадцяти приладів була випробувана щодо терміну безвідмовної роботи кожного з них t_i. Результати випробувань наведено у вигляді дискретного статистичного розподілу.

t_i	100	170	240	310	380
n_i	2	5	10	2	1

На рівні значущості $\alpha = 0{,}01$ побудувати довірчий інтервал для середнього часу безвідмовної роботи приладу.

Послідовність рішення.

1. Використовуючи матеріал лабораторної роботи 4, знаходимо середнє вибіркове та дисперсію D_B.

2. Для точності виконання розрахунків, використовуємо виправлене середнє квадратичне відхилення, що дорівнює $s_x = \sqrt{\frac{n}{n-1} D_B}$.

3. Для розрахунку параметра розподілу використовуємо вбудовану функцію =T.INV.2T(рівень значущості; ступінь свободи), що повертає двобічний обернений t-розподіл Стьюдента. У нашому випадку отримане значення рівне 2,86. Перевірити точність можна за таблицями для розподілу Стьюдента.

4. Розмах довірчого інтервалу знаходимо, використовуючи вбудовану функцію =CONFIDENCE.T(альфа;станд_відхилення;розмір). Отримане значення дорівнює 43,28.

5. Межі довірчого інтервалу знаходимо як різницю та суму середнього вибіркового і знайденого розмаху.

Отже, з надійністю 99% можна стверджувати, що середнє для генеральної сукупності буде знаходитися в інтервалі від 179,22 до 265,78.

Виконання розрахунків проілюстрвано на рисунку нижче.

	A	B	C	D	E	F	G
1	**ti**	**ni**	**ti*ni**	**ni*ti^2**		**Розподіл Стьюдента**	=T.INV.2T(D9;D10)
2	100	2	=A2*B2	=B2*A2^2		**Розмах довірчого інтералу**	=CONFIDENCE.T(D9;B12;B7)
3	170	5	=A3*B3	=B3*A3^2		**Межа 1**	=B9-G2
4	240	10	=A4*B4	=B4*A4^2		**Межа 2**	=B9+G2
5	310	2	=A5*B5	=B5*A5^2			
6	380	1	=A6*B6	=B6*A6^2			
7		**=SUM(B2:B6)**	**=SUM(C2:C6)**	**=SUM(D2:D6)**			
8							
9	**Середнє:**	=C7/B7	**α**	0,01			
10	**Квадрат середнього:**	=B9^2	**Ступені вільності**	=B7-1			
11	**Дисперсія:**	=D7/B7-B10					
12	**Стандартне відхилення:**	=SQRT(B7/(B7-1)*B11)					

Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи.

1. Знайти при $\alpha = 0{,}05$ довірчий інтервал оцінки середнього генеральної сукупності, якщо відомі вибіркова середня, об'єм вибірки та середнє квадратичне відхилення генеральної сукупності.

Варіант	$\overline{X}$	n	s_x
1	14	25	5
2	10,2	16	4
3	16,8	25	5
4	2000	1600	40
5	15	40	0,3
6	101	30	5
7	256	150	4,56
8	1,05	31	0,01
9	2,04	37	0,12
10	390	220	8,8

2. Випадково вибрана партія з двадцяти приладів була випробувана щодо терміну безвідмовної роботи кожного з них t_i. Результати випробувань наведено у вигляді дискретного статистичного розподілу. На рівні значущості $\alpha = 0{,}01$ побудувати довірчий інтервал для середнього часу безвідмовної роботи приладу.

Варіант	Завдання					
1	t_i	2	5	7	8	10
	n_i	6	4	5	5	5
2	t_i	-2	0	3	5	7
	n_i	7	13	14	15	11

3	t_i	-2	-1	0	1	2
	n_i	1	3	2	4	2
4	t_i	2	3	5	6	8
	n_i	15	10	25	30	35
5	t_i	2	4	6	8	10
	n_i	2	3	5	1	4
6	t_i	6	8	10	12	14
	n_i	5	1	4	2	3
7	t_i	3	5	7	10	15
	n_i	2	4	7	4	3
8	t_i	-2	-1	0	1	2
	n_i	7	13	14	15	11
9	t_i	2	3	5	6	7
	n_i	6	4	5	5	3
10	t_i	4	8	12	16	20
	n_i	1	3	5	4	4

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 6
ПЕРЕВІРКА СТАТИСТИЧНОЇ ГІПОТЕЗИ ПРО ВИГЛЯД ЗАКОНУ РОЗПОДІЛУ ДОСЛІДЖУВАНОЇ ВЕЛИЧИНИ

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для перевірки статистичних гіпотез.

Основні поняття: нульова гіпотеза, альтернативна гіпотеза, критерії асиметрії та ексцесу, критерій згоди Пірсона.

Теоретичні відомості та хід виконання роботи

Перевірка гіпотези про вигляд закону розподілу досліджуваної величини має велике значення для прикладних досліджень. Необхідність такої перевірки виникає при виборі критерію, оскільки для багатьох з них висувається вимога нормального розподілу статистичних даних.

Припустимо, що з деякої генеральної сукупності X, яка розглядається як випадкова величина, обрана вибірка $\{x_1, x_2, \dots, x_n\}$. За даними вибірки побудовано статистичний ряд (табл. 1), що містить варіанти x_i та відповідні частоти n_i, $i = \overline{1,k}$, де k – кількість варіант у випадку дискретного ряду. У випадку інтервального ряду x_i – середини інтервалів, k – кількість інтервалів.

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

Отриманий на основі вибіркових даних статистичний ряд називається емпіричним законом розподілу величини X. За даними статистичного ряду можна знайти числові характеристики, які є вибірковими параметрами закону розподілу X. Вид закону розподілу визначається відповідно до умов формування вибірки або залежно від виду графіка емпіричної густини розподілу (гістограми) у випадку неперервної випадкової величини X і полігону частот, якщо величина X дискретна. Параметри обраного закону розподілу змінюються відповідними вибірковими параметрами.

Закон розподілу випадкової величини X, параметрами якого є відповідні вибіркові числові характеристики, називається теоретичним законом розподілу.

При здійсненні такої заміни немає впевненості, що закон розподілу обраний правильно. Тому розроблено процедуру, яка дозволяє оцінити степінь відповідності обраного закону даним вибірки.

Спосіб 1. Критерії асиметрії та ексцесу

Критерії асиметрії та ексцесу застосовують для приблизної перевірки гіпотези про нормальність емпіричного розподілу. Асиметрія характеризує ступінь несиметричності, а ексцес – ступінь загостреності (згладженості) кривої диференціальної функції емпіричного розподілу в порівнянні з функцією густини нормального розподілу.

Для нормального розподілу $\mathrm{N}(\mu, \sigma)$ з середнім μ і дисперсією σ^2 третій і четвертий моменти мають сенс асиметрії і ексцесу та дорівнюють нулю.

$$A = \frac{m_3}{\sigma^3} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^3 p_i = 0,$$

$$E = \frac{m_4}{\sigma^4} - 3 = \left[\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^4 p_i\right] - 3 = 0.$$

Дисперсії асиметрії та ексцесу відповідно дорівнюють:

$$\mathrm{D}(A) = \frac{6\cdot(n-1)}{(n+1)\cdot(n+3)},\ \mathrm{D}(E) = \frac{24\cdot n\cdot(n-2)\cdot(n-3)\cdot(n-5)}{(n-1)^2\cdot(n+1)\cdot(n+3)}.$$

Емпіричний розподіл вважається близьким до нормального (приймається нульова гіпотеза), якщо виконуються умови:

$$|A_x| \le 3\sqrt{D(A)} \text{ і } |E_x| \le 5\sqrt{D(E)}.$$

Технологічно у цьому методі розраховуються показники

$$t_A = \frac{|A_x|}{\sqrt{D(A)}} \text{ і } t_E = \frac{|E_x|}{\sqrt{D(E)}}$$

Про достовірну відмінність емпіричного розподілу від нормального свідчать показники t_A і t_E, значення яких 3 і більше.

Приклад 1.

Емпіричні дані		Критерії t_A і t_E	
j	xi	Розрахунки	
1	4		
2	4	n	18
3	4	A	-0,23884
4	5	E	-1,53198
5	6	D(A)	0,505608
6	7	D(E)	3,107397
7	8	t_A	0,472385
8	9	t_E	0,49301
9	10		
10	11	$X_{сер}$	9,944444
11	12	s_x	4,151148
12	13		
13	13		
14	14		
15	14		
16	15		
17	15		
18	15		

ипіричні да		Критерії t_A і t_E	
j	xi	Розрахунки	
1	4		
2	4	n	=COUNT(B3:B20)
3	4	A	=SKEW(B3:B20)
4	5	E	=KURT(B3:B20)
5	6	D(A)	=SQRT(6*(D4-1)/(D4+1)/(D4+3))
6	7	D(E)	=SQRT((24*D4*(D4-2)*(D4-3)*(D4-5)/(D4-1)^2/(D4+3)/(D4+5)))
7	8	t_A	=ABS(D5)/D7
8	9	t_E	=ABS(D6)/D8
9	10		
10	11	$X_{сер}$	=AVERAGE(B3:B20)
11	12	s_x	=STDEV.S(B3:B20)
12	13		
13	13		
14	14		
15	14		
16	15		
17	15		
18	15		

Спосіб 2. Критерій згоди χ^2

Критерії здійснення такої перевірки називаються критеріями згоди, найбільш відомим з яких є критерій Пірсона χ^2. Для емпіричних даних він розраховується за формулою:

$$\chi^2_{\text{емп}} = \sum_{i=1}^{k}\frac{(m_i - np_i)^2}{np_i}$$

де m_i – кількість значень, що потрапляють в i-й інтервал, n – обсяг вибірки, p_i – теоретична ймовірність величини потрапити в i-й інтервал).

У ході перевірки формулюються наступні гіпотези:

H_0: емпіричний закон ***не відрізняється*** від нормального;

H_1: емпіричний закон ***відрізняється*** від нормального.

Послідовність розрахунку критерію в MS Excel та відповідні формули наведено нижче на рисунках. Дані для розрахунків записано у стовпчику B. Кількість класів k розраховується за формулою Стерджеса (комірка D12), λ – ширина інтервалу класу.

При розрахунках вважаються відомими генеральне середнє $\mu = 10$ і генеральне стандартне відхилення $\sigma = 4$.

A	B	C	D	E	F
Емпіричні дан					
j	**xi**	**Інтервали**	**x_i**	**x_{i+1}**	**m_i**
1	4	1	−∞	4	=FREQUENCY(B3:B20;D3:D8)
2	4	2	4	6	
3	4	3	6	8	
4	5	4	8	10	
5	6	5	10	12	
6	7	6	12	+∞	
7	8	Суми:			=SUM(F3:F8)
8	9				
9	10				
10	11	**k**	=ROUND(1+3,32*LOG(COUNT(B3:B20));0)		
11	12	**λ**	=ROUND((MAX(B3:B20)-MIN(B3:B20))/D12;0)		
12	13	**μ**	10		
13	13	**σ**	4		
14	14				
15	14				
16	15				
17	15				
18	15				

G	H	I	J	K	L
$F(x_i)$	**$F(x_{i+1})$**	**p_i**	**np_i**	**$(m_i-np_i)^2$**	**(mi-npi)2/npi**
=0	=NORM.DIST(E3;$D$14;$D$15;1)	=H3-G3	=$F$9*I3	=(F3-J3)^2	=K3/J3
=NORM.DIST(D4;$D$14;$D$15;1)	=NORM.DIST(E4;$D$14;$D$15;1)	=H4-G4	=$F$9*I4	=(F4-J4)^2	=K4/J4
=NORM.DIST(D5;$D$14;$D$15;1)	=NORM.DIST(E5;$D$14;$D$15;1)	=H5-G5	=$F$9*I5	=(F5-J5)^2	=K5/J5
=NORM.DIST(D6;$D$14;$D$15;1)	=NORM.DIST(E6;$D$14;$D$15;1)	=H6-G6	=$F$9*I6	=(F6-J6)^2	=K6/J6
=NORM.DIST(D7;$D$14;$D$15;1)	=NORM.DIST(E7;$D$14;$D$15;1)	=H7-G7	=$F$9*I7	=(F7-J7)^2	=K7/J7
=NORM.DIST(D8;$D$14;$D$15;1)	1	=H8-G8	=$F$9*I8	=(F8-J8)^2	=K8/J8
		=SUM(I3:I8)	=SUM(J3:J8)		=SUM(L3:L8)

χ^2	=SUM(L3:L8)
$\chi^2_{0.1}$	=CHISQ.INV(0,9;6-1)
$\chi^2_{0.05}$	=CHISQ.INV(0,95;6-1)

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Емпіричні дані**											
2	**j**	**xi**	**терва**	**x_i**	**x_{i+1}**	**m_i**	**$F(x_i)$**	**$F(x_{i+1})$**	**p_i**	**np_i**	**$(m_i-np_i)^2$**	**(mi-npi)2/npi**
3	1	4	1	−∞	4	3	0,000	0,067	0,067	1,203	3,231	2,687
4	2	4	2	4	6	2	0,067	0,159	0,092	1,653	0,120	0,073
5	3	4	3	6	8	2	0,159	0,309	0,150	2,698	0,487	0,181
6	4	5	4	8	10	2	0,309	0,500	0,191	3,446	2,092	0,607
7	5	6	5	10	12	2	0,500	0,691	0,191	3,446	2,092	0,607
8	6	7	6	12	+∞	7	0,691	1,000	0,309	5,554	2,092	0,377
9	7	8	Суми:			18			1	18		4,530622268
10	8	9										
11	9	10										
12	10	11	**k**	5			χ^2	4,5306				
13	11	12	**λ**	2			$\chi^2_{0.1}$	9,2364				
14	12	13	**μ**	10			$\chi^2_{0.05}$	11,07				
15	13	13	**σ**	4								
16	14	14										
17	15	14										
18	16	15										
19	17	15										
20	18	15										

Теоретичне значення критерію $\chi^2_{\text{теор}}$ розраховуємо з використанням функції CHISQ.INV() для довірчої ймовірності 0,9 (рівень значущості 0,1) та довірчої ймовірності 0,95 (рівень значущості 0,05).

Оскільки розраховане значення $\chi^2_{\text{емп}} \approx 4{,}53$ не перевищує критичного значення навіть на рівні значущості 0,1 $\chi^2_{\text{теор}} \approx 9{,}24$, то приймається нульова гіпотеза про відповідність розподілу нормальному.

Висновок: розбіжності емпіричного і теоретичного нормального розподілів можуть мати винятково випадковий характер.

У випадку, якщо дані, для яких перевіряється гіпотеза про нормальність розподілу, згруповані за інтервалами, для розрахунку середнього значення та дисперсії потрібно використовувати середнє за інтервалами.

1. Середини розрядів: $x_i = \frac{z_{i-1}+z_i}{2}$ (z_{i-1} і z_i – ліва й права межі інтервалів).

2. Середнє для вибірки: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot n_i$.

3. Виправлена дисперсія та стандартне відхилення: $D_{\text{В}} = \frac{1}{n-1}(\sum_{i=1}^{n} n_i \cdot x_i^2 - n \cdot \bar{x}^2)$, $s_x = \sqrt{D_B}$.

Приклад і результат розрахунків наведено нижче.

	A	B	C	D	E	F	G	H	I
1	**Ліва межа**	**Права межа**	**Середнє інтервалу**	**ni**	**xi*ni**	**ni*xi^2**			
2	15	19	=(A2+B2)/2	45	=D2*C2	=D2*C2^2		**Середнє**	=E7/D7
3	19	23	=(A3+B3)/2	125	=D3*C3	=D3*C3^2		**n*x_avg^2**	=D7*I2^2
4	23	27	=(A4+B4)/2	175	=D4*C4	=D4*C4^2		**Дисперсія**	=(F7-I3)/(D7-1)
5	27	31	=(A5+B5)/2	115	=D5*C5	=D5*C5^2		**Стандартне відхилення**	=SQRT(I4)
6	31	35	=(A6+B6)/2	40	=D6*C6	=D6*C6^2			
7				**=SUM(D2:D6)**	**=SUM(E2:E6)**	**=SUM(F2:F6)**			

	A	B	C	D	E	F	G	H	I
1	**Ліва межа**	**Права межа**	**Середнє інтервалу**	**ni**	**xi*ni**	**ni*xi^2**			
2	15	19	17	45	765	13005		**Середнє**	24,84
3	19	23	21	125	2625	55125		**n*x_avg^2**	308512,80
4	23	27	25	175	4375	109375		**Дисперсія**	18,57
5	27	31	29	115	3335	96715		**Стандартне відхилення**	4,31
6	31	35	33	40	1320	43560			
7				**500**	**12420**	**317780**			

Довірчі інтервали розраховуються з урахуванням певних вимог до генеральної сукупності (вимога **нормальності розподілу даних**).

Для нормального розподілу модель інтервальної оцінки середнього генеральної сукупності μ має вигляд:

$$\mu \in [\bar{X} - \Delta, \bar{X} + \Delta]$$

де $\Delta = \frac{z_{\alpha/2} \cdot s_x}{\sqrt{n}}$, $\bar{X}$ і s_x – вибіркове середнє і стандартне відхилення, n – обсяг вибірки, $z_{\alpha/2}$ – параметр стандартного нормального розподілу, α – рівень значущості (імовірність того, що відхилення вибіркового від генерального середнього не перевищить Δ за абсолютним значенням.

Завдання для самостійного виконання

Нехай є вибірка деякої випадкової величини X у вигляді інтервального статистичного ряду. Потрібно:

а) побудувати гістограму та графік емпіричної функції розподілу випадкової величини;

б) у припущенні нормального закону розподілу випадкової величини X знайти довірчий інтервал для математичного сподівання з довірчою ймовірністю 0,95 ;

в) за критерієм згоди Пірсона перевірити гіпотезу про нормальний розподіл випадкової величини X для рівня значущості α =0,01.

Завдання виконується за варіантами, що відповідають списку групи.

	Завдання						
1	Інтервал	(21;23)	(23;25)	(25;27)	(27;29)	(29;31)	(31;33)
	Частота	30	70	65	30	25	5
2	Інтервал	(40;45)	(45;50)	(50;55)	(55;60)	(60;65)	(65;70)
	Частота	50	100	105	40	35	10
3	Інтервал	(100;105)	(105;110)	(110;115)	(115;120)	(120;125)	(125;130)
	Частота	45	105	100	40	10	2
4	Інтервал	(10;15)	(15;20)	(20;25)	(25;30)	(30;35)	(35;40)
	Частота	60	140	135	55	20	4
5	Інтервал	(3;8)	(8;13)	(13;18)	(18;23)	(23;28)	(28;33)
	Частота	6	8	15	40	16	8
6	Інтервал	(1;3)	(3;5)	(5;7)	(7;9)	(9;11)	(11;13)
	Частота	2	4	6	10	18	20
7	Інтервал	(6;16)	(16;26)	(26;36)	(36;46)	(46;56)	(56;66)
	Частота	8	7	16	35	15	8
8	Інтервал	(5;10)	(10;15)	(15;20)	(20;25)	(25;30)	(30;35)
	Частота	7	8	15	18	23	19
9	Інтервал	(-20;-10)	(-10;0)	(0;10)	(10;20)	(20;30)	(30;40)
	Частота	20	47	80	89	40	16
10	Інтервал	(3;7)	(7;11)	(11;15)	(15;19)	(19;23)	(23;27)
	Частота	6	16	19	17	15	14

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 7
ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ ПРО РІВНІСТЬ ПАРАМЕТРІВ

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для перевірки статистичних гіпотез.

Основні поняття: нульова гіпотеза, альтернативна гіпотеза, критерії асиметрії та ексцесу, критерій Стьюдента, критерій Крамера-Велча, критерій Фішера.

Теоретичні відомості та хід виконання роботи

Під час досліджень доволі часто виникає необхідність перевірити чи розрізняються генеральні сукупності, з яких узято вибірки. Наприклад, чи відрізняються між собою експериментальна і контрольна групи студентів за результатами тестування академічних досягнень. Методи перевірки статистичних гіпотез про однорідність вибірок можуть реалізовуватися з використанням параметричних та непараметричних критеріїв для незалежних (незв'язаних) і залежних (зв'язаних) двох і більше вибірок.

Для варіанту незалежних вибірок постановка математично-статистичної задачі формулюється наступним чином: дві вибірки обсягом n_1 і n_2 взято випадковим методом з двох генеральних сукупностей, неперервні функції розподілу $F_1(x)$ і $F_2(x)$ яких є невідомими. Потрібно перевірити їх однорідність (неоднорідність).

Нульова й альтернативна гіпотези мають вигляд:

$$H_0: F_1(x) = F_2(x)$$
$$H_1: F_1(x) \neq F_2(x)$$

У математичній статистиці використовуються декілька методів перевірки однорідності, тому важливим є уміння обрати оптимальний для дослідницької задачі.

Критерій Стьюдента t. Вираз критерію Стьюдента має вигляд:

$$t_{\text{емп}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (1)$$

де $\bar{X}_1$ і $\bar{X}_2$, s_1^2 і s_2^2, n_1 і n_2 – середні значення, дисперсії та обсяги першої і другої вибірок, відповідно.

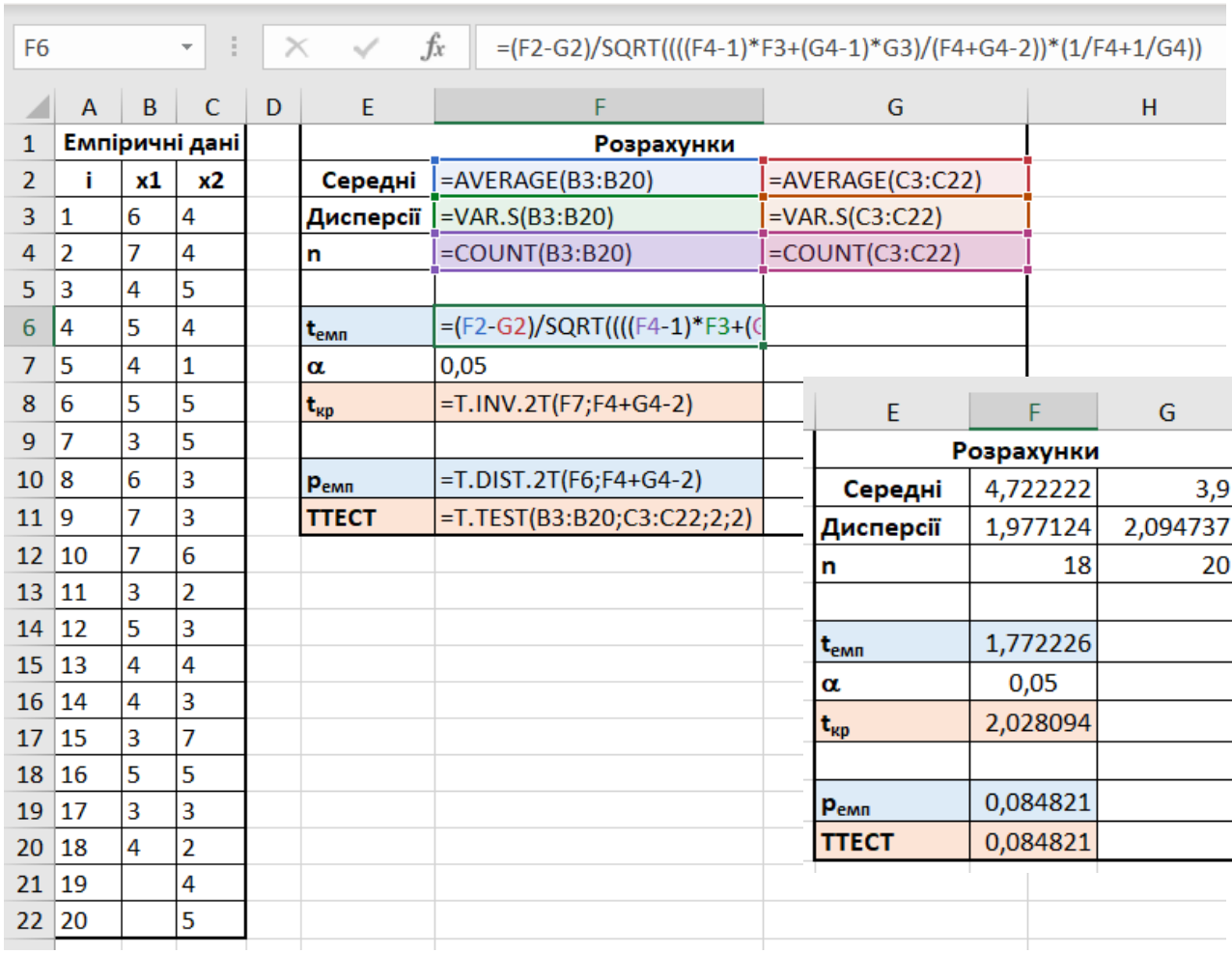
Критичне значення критерію $t_{\text{кр}}$ для заданого рівня значущості та числа ступенів вільності $n_1 + n_2 - 2$ можна отримати з таблиць розподілу Стьюдента або за допомогою функції =TINV(імовірність; кількість ступенів вільності).

Приклад 1. Перевірити статистичні гіпотези на рівні значущості 0,05 щодо однорідності двох незалежних вибірок за критерієм Стьюдента (дані та розрахунки наведено на рисунках).

Досліджуваній ситуації відповідає варіант неспрямованих гіпотез:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

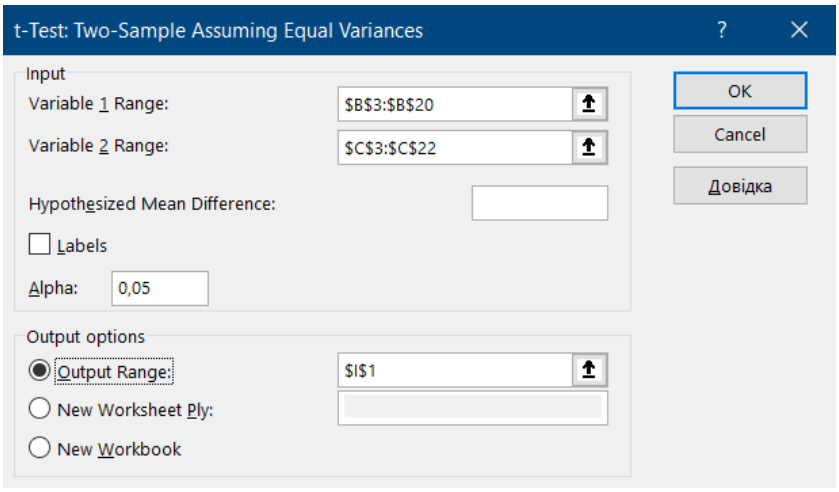
Емпіричний критерій розраховується за формулою (1), а теоретичне критичне значення – з використанням вбудованої функції MS Excel =T.INV.2T(), яка повертає значення для двобічного критерію. Якщо абсолютна величина емпіричного значення $t_{\text{емп}}$ менше від критичного $t_{\text{кр}}$, то нульова гіпотеза приймається на рівні значущості 0,05.



Зазначимо, що статистика критерію Стьюдента перевіряє не збіг функцій розподілу вибірок, а збіг характеристик випадкових величин – математичних сподівань (середніх).

Перевірку гіпотез можна здійснити також шляхом визначення ймовірності $p_{\text{емп}}$ для розрахованого значення $t_{\text{емп}}$: =T.DIST.2T(критерій; число ступенів вільності). Якщо $p_{\text{емп}} \leq \alpha$, то нульова гіпотеза відхиляється. Зручним способом перевірки гіпотези однорідності вибірок є використання вбудованої функції T.TEST(перша вибірка; друга вибірка; боки; тип) та порівняння отриманого значення з рівнем значущості.

У MS Excel для перевірки гіпотези можна скористатися пакетом Data Analysis, функція t-Test: Two-Sample Assuming Equal Variances. Приклад розрахунків наведено нижче.



t-Test: Two-Sample Assuming Equal Variances		
	Variable 1	*Variable 2*
Mean	4,722222222	3,9
Variance	1,977124183	2,094736842
Observations	18	20
Pooled Variance	2,039197531	
Hypothesized Me	0	
df	36	
t Stat	1,772225634	
P(T<=t) one-tail	0,042410392	
t Critical one-tail	1,688297714	
P(T<=t) two-tail	0,084820784	
t Critical two-tail	2,028094001	

Критерій Крамера-Велча Т. Вираз критерію Крамера-Велча має вигляд:

$$T_{\text{емп}} = \frac{\sqrt{n_1 n_2}(\bar{X}_1 - \bar{X}_2)}{\sqrt{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}} \quad (2)$$

де $\bar{X}_1$ і $\bar{X}_2$, s_1^2 і s_2^2, n_1 і n_2 – середні значення, дисперсії та обсяги першої і другої вибірок, відповідно. Невідомі для генеральних сукупностей дисперсії замінюються вибірковими значеннями. При зростанні обсягів вибірок розподіл статистики Крамера-Велча Т збігається до стандартного нормального розподілу.

Правило ухвалення рішень для критерію Крамера-Велча є наступним: якщо абсолютне значення емпіричного критерію менше від значення, розрахованого з використанням нормального розподілу, то гіпотеза однорідності (рівності математичних сподівань) приймається на рівні значущості 0,05:

$$|T_{\text{емп}}| < z\left(1 - \frac{\alpha}{2}\right).$$

Нижче наведено виконання розрахунків для перевірки критерію Крамера-Велча в табличному процесорі.

	E	F	G
	Розрахунки		
Середні	4,72222222	3,9	
Дисперсії	1,97712418	2,094737	
n	18	20	
$T_{емп}$	1,77230036		
α	0,05		
$t_{кр}$	1,95996398		

	E	F	G
	Розрахунки		
Середні	=AVERAGE(B3:B20)	=AVERAGE(C3:C22)	
Дисперсії	=VAR.S(B3:B20)	=VAR.S(C3:C22)	
n	=COUNT(B3:B20)	=COUNT(C3:C22)	
$T_{емп}$	=SQRT(F4*G4)*(F2-G2)/SQRT(F4*F3+G4*G3)		
α	0,05		
$t_{кр}$	=NORM.S.INV(1-F7/2)		

Методи Стьюдента й Крамера-Велча обмежуються перевіркою рівності математичних сподівань або інших параметрів розподілу. Тому, для порівняння власне розподілів використовують непараметричні методи, які будуть розглядатися в наступній роботі.

Критерій Фішера F. Для перевірки гіпотези про рівність дисперсій двох незалежних генеральних сукупностей використовується критерій Фішера F:

$$F_{\text{емп}} = \frac{s_1^2}{s_2^2}$$

де s_1^2 і s_2^2 – дисперсії вибірок. Обсяги вибірок можуть бути однаковими або різними. Емпіричне значення критерію порівнюється з теоретичним для заданого рівня значущості. На початку перевірки формулюються гіпотези у припущенні, що досліджуваний параметр має нормальний розподіл, вибірки незв'язані.

$$H_0: \sigma_1 = \sigma_2$$
$$H_1: \sigma_1 \neq \sigma_2$$

Приклад розрахунків наведено нижче.

	A	B	C	D	E	F	G
1	**Емпіричні дані**				**Розрахунки**		
2	**i**	**x1**	**x2**		**Середні**	=AVERAGE(B3:B16)	=AVERAGE(C3:C18)
3	1	6	6		**Дисперсії**	=VAR.S(B3:B16)	=VAR.S(C3:C18)
4	2	2	4		**n**	=COUNT(B3:B16)	=COUNT(C3:C18)
5	3	4	4				
6	4	4	6		**$F_{емп}$**	=F3/G3	
7	5	3	5		**α**	0,05	
8	6	4	4		**F.INV**	=F.INV(F7/2;F4-1;G4-1)	
9	7	6	5		**F.INV.RT**	=F.INV.RT(F7/2;F4-1;G4-1)	
10	8	5	5				
11	9	2	4		**FTEST**	=F.TEST(B3:B16;C3:C18)/2	
12	10	4	6				
13	11	4	4				
14	12	3	4				
15	13	4	5				
16	14	5	5				
17	15		6				
18	16		4				

E	F	G
Розрахунки		
Середні	4	4,8125
Дисперсії	1,5384615	0,69583
n	14	16
$F_{емп}$	2,2109627	
α	0,05	
F.INV	0,3275774	
F.INV.RT	2,9249044	
FTEST	0,0718252	

Для двобічної моделі ми визначаємо теоретичні значення критерію і зліва F.INV і справа F.INV.RT. Оскільки емпіричне значення становить 2,21 і знаходиться в інтервалі від 0,327 і 2,924, то нульова гіпотеза може бути прийнята. Іншим способом є використання функції F.TEST(перша вибірка; друга вибірка). Отримане значення ділиться пополам. Воно становить 7,18% і більше від необхідної точності 5%.

Для розрахунків із використанням пакету Data Analysisвикористовуємо функцію F-Test Two-Sample for Variances.

F-Test Two-Sample for Variances		
	Variable 1	*Variable 2*
Mean	4	4,8125
Variance	1,538461538	0,695833333
Observations	14	16
df	13	15
F	2,21096269	
P(F<=f) one-tail	0,071825197	
F Critical one-tail	2,44811021	

Завдання для самостійного виконання

Нехай є дві вибірки деяких випадкових величин, для яких відомі об’єми та знайдені вибіркові середні та вибіркові дисперсії.

Використовуючи методи Стьюдента і Крамера-Велча перевірити гіпотезу про рівність математичних сподівань вибірок при рівні значущості 0,01. Для вибірок здійснити також перевірку гіпотези про рівність дисперсій за критерієм Фішера.

Завдання виконується за варіантами, що відповідають списку групи.

Варіант	n	$\overline{X}$	s_x^2	m	$\overline{Y}$	s_y^2
1	10	14,3	1,7	8	15,3	2,2
2	12	31,2	0,84	18	29,2	0,40
3	145	31,4	3,36	200	28,84	3,51
4	25	2,3	0,25	26	2,48	0,108
5	14	36,2	2,67	12	35,1	2,55
6	50	1282	80	60	1208	94
7	16	10	1,1	36	4	1,4
8	16	37,5	1,21	25	36,8	1,44
9	40	84	10,1	54	77,5	8,4
10	50	85	100	70	78	74

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 8

ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ ПРО РІВНІСТЬ ПАРАМЕТРІВ З ВИКОРИСТАННЯМ НЕПАРАМЕТРИЧНИХ КРИТЕРІЇВ

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для перевірки статистичних гіпотез.

Основні поняття: нульова гіпотеза, альтернативна гіпотеза, критерій Колмогорова-Смирнова, критерій Вілкоксона-Манна-Вітні, критерій Лемана-Розенблатта.

Теоретичні відомості та хід виконання роботи

Критерій Колмогорова-Смирнова λ . Вираз критерію **Колмогорова-Смирнова** має вигляд:

$$\lambda_{\text{емп}} = max|F_1(x) - F_2(x)| \quad (1)$$

де $F_1(x)$, $F_2(x)$ – емпіричні функції розподілу вибірок обсягом n_1 і n_2.

Критерій дозволяє знайти точку, у якій сума накопичених розбіжностей між двома розподілами $F_1(x)$ і $F_2(x)$ є найбільшою і оцінити достовірність цієї розбіжності. Для λ-критерію зіставляють накопичені (інтегральні) частоти. Нульова гіпотеза H_0 свідчить про те, що відмінності між двома розподілами недостовірні.

Приклад 1. Зробити статистичні висновки на рівні значущості 0,05 щодо однорідності двох незалежних вибірок за критерієм Колмогорова-Смирнова (дані та розрахунки наведено на рисунках).

Досліджуваній ситуації відповідає варіант неспрямованих гіпотез:

H_0 : відмінності між двома розподілами *недостовірні* (судячи з точки максимальної накопиченої розбіжності між ними);

H_1: відмінності між двома розподілами *достовірні* (судячи з точки максимальної накопиченої розбіжності між ними).

Нижче наведено виконання розрахунків.

	A	B	C	D	E	F	G	H	I	J
1	Емпіричні дані				Варіанти	Частоти				$\lvert F_1-F_2\rvert$
2	i	x	y			диференціальні		інтегральні		
3	1	6	4			n_1	n_2	F_1	F_2	
4	2	7	4		0	=FREQUENCY(B3:B20;E4:E11)	=FREQUE	=F4/$F$13	=G4/$G$13	=ABS(H4-I4)
5	3	4	5		1			=F5/$F$13+H4	=G5/$G$13+I4	=ABS(H5-I5)
6	4	5	4		2			=F6/$F$13+H5	=G6/$G$13+I5	=ABS(H6-I6)
7	5	4	1		3			=F7/$F$13+H6	=G7/$G$13+I6	=ABS(H7-I7)
8	6	5	5		4			=F8/$F$13+H7	=G8/$G$13+I7	=ABS(H8-I8)
9	7	3	5		5			=F9/$F$13+H8	=G9/$G$13+I8	=ABS(H9-I9)
10	8	6	3		6			=F10/$F$13+H9	=G10/$G$13+I	=ABS(H10-I10)
11	9	7	3		7			=F11/$F$13+H10	=G11/$G$13+I	=ABS(H11-I11)
12	10	3	6							
13	11	7	2		Суми	=SUM(F4:F11)				=MAX(J4:J11)
14	12	3	3							0,294
15	13	5	4							
16	14	4	3							
17	15	4	7							
18	16	3	5							
19	17	5	3							
20	18	4	2							
21	19		4							
22	20		5							

E	F	G	H	I	J
Варіанти	Частоти				$\lvert F_1-F_2\rvert$
	диференціальні		інтегральні		
	n_1	n_2	F_1	F_2	
0	0	0	0,00	0,00	0,000
1	0	1	0,00	0,05	0,050
2	0	2	0,00	0,15	0,150
3	4	5	0,22	0,40	0,178
4	5	5	0,50	0,65	0,150
5	4	5	0,72	0,90	0,178
6	2	1	0,83	0,95	0,117
7	3	1	1,00	1,00	0,000
	0	0			
Суми	18	20		$\lambda_{емп}$	0,178
				$\lambda_{0,05}$	0,294

Оскільки емпіричне значення критерію $\lambda_{\text{емп}}$ (взяте з довідкових таблиць) менше від критичного значення, гіпотеза H_0 приймається на рівні значущості 0,05. Таким чином, на рівні значущості 0,05 відсутні підстави говорити про неоднорідність незалежних вибірок.

Критерій Вілкоксона-Манна-Вітні U. Статистика U-критерію визначається наступним чином. Усі X-елементи першої та Y-елементи другої вибірок об'єднуються. Об'єднана вибірка $x_1, x_2, \dots, x_{n1}, y_1, y_2, \dots, y_{n2}$ упорядковується за зростанням (n_1 і n_2 – обсяги вибірок). Далі аналізуються ранги – позиції, які в об'єднаному варіаційному ряді займають елементи першої вибірки. Їх сума є статистикою Вілкоксона:

$$T_x = R_1 + R_2 + \cdots + R_{n1}$$

Статистика Манна-Вітні U визначається формулою:

$$U_{\text{емп}} = n_1 \cdot n_2 + \frac{n_x \cdot (n_x+1)}{2} - T_x$$

Оскільки T_x і U лінійно зв'язані, то мова часто йдеться не про два критерії – Вілкоксона і Манні-Вітні, а про один об'єднаний – Вілкоксона-Манна-Вітні. За допомогою критерію визначається зона значень між двома чисельними рядами, що перехрещуються. Чим менше емпіричне значення критерію $U_{\text{емп}}$, тим більш імовірно, що відмінності достовірні. Коли обсяги вибірок нескінченно зростають, розподіли статистик Вілкоксона і Манні-Вітні є асимптотично нормальними.

Приклад 2. Проаналізуємо вибірки з попереднього прикладу з використанням критерію Вілкоксона-Манна-Вітні U.

Досліджуваній ситуації відповідає варіант гіпотез:

H_0: відмінності між показниками ознаки не є статистично значущими;

H_1: відмінності між показниками ознаки є статистично значущими.

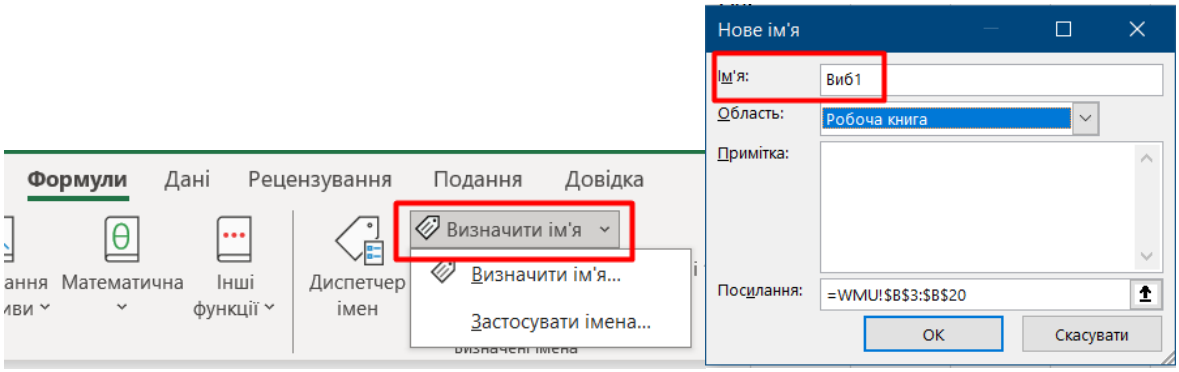
	A	B	C	D	E	F	G	H
1	**Емпіричні дані**			**Ранги**			**Розрахунки**	
2	**i**	**x**	**y**	**Ранг 1**	**Ранг 2**			
3	1	6	4	33	17,5		**n_1**	18
4	2	7	4	36,5	17,5		**n_2**	20
5	3	4	5	17,5	27		**T_1**	403
6	4	5	4	27	17,5		**T_2**	338
7	5	4	1	17,5	1		**n_x**	18
8	6	5	5	27	27		**T_x**	403
9	7	3	5	8	27		**$U_{емп}$**	128
10	8	6	3	33	8		**$U_{0,05}$**	123
11	9	7	3	36,5	8			
12	10	3	6	8	33			
13	11	7	2	36,5	2,5			
14	12	3	3	8	8			
15	13	5	4	27	17,5			
16	14	4	3	17,5	8			
17	15	4	7	17,5	36,5			
18	16	3	5	8	27			
19	17	5	3	27	8			
20	18	4	2	17,5	2,5			
21	19		4		17,5			
22	20		5		27			

G	H
Розрахунки	
n_1	=COUNT(Виб1)
n_2	=COUNT(Виб2)
T_1	=SUM(D3:D20)
T_2	=SUM(E3:E22)
n_x	=IF(H5>H6;H3;H4)
T_x	=IF(H5>H6;H5;H6)
$U_{емп}$	=H3*H4+H7*(H7+1)/2-H8
$U_{0,05}$	123

Для розрахунку рангів (комірки стовпчиків D і E) використовується формула MS Excel за прикладом, наведеним нижче:

=(COUNT(Виб1:Виб2)+1-RANK.AVG(B3;Виб1:Виб2;1)-
RANK.AVG(B3;Виб1:Виб2;0))/2+RANK.AVG(B3;Виб1:Виб2;1)

Виб1 і Виб2 – змінні, введені для роботи з вибірками. Для створення Виб1 потрібно виділити мишкою комірки B3:B20, перейти до меню Формули і обрати опцію Визначити ім'я. Далі, у вікні діалогу написати назву, що ми присвоюємо вибірці. Посилання, яке в подальшому автоматично використовуватиметься для цього масиву, включає назву аркуша та адреси виділених комірок. Для Виб2 дії повторюються.



Емпіричні значення критеріїв T_1 і T_2 – це суми рангів. Для оцінки гіпотези потрібно обрати більше зі значень та відповідний обсяг вибірки. У нашому випадку це T_1 і n_1. Оскільки $U_{\text{емп}} > U_{0,05}$, нульова гіпотеза приймається на рівні значущості 0,05. Це дозволяє стверджувати, що відмінності в показниках ознаки не є статистично значущими.

Завдання для самостійного виконання

Використовуючи критерії Колмогорова-Смирнова та Вілкоксона-Манні-Вітні, оцінити відмінності у показниках на рівні значущості 0,05 за даними таблиці.

Завдання виконується за варіантами, що відповідають списку групи.

1.

X	1	7	14	3	1	6	7	10	4	9	11	5	3	10	9	12
Y	5	6	10	7	3	4	9	8	1	11	13	2	8	1	2	4

2.

X	1	3	11	28	91	110	95	54	45	12	3
Y	3	25	23	45	112	88	87	37	26	6	1

3.

X	3	4	4	5	5	5	5	5	6	6	7	4	4	6	3
Y	3	3	3	4	4	4	4	4	5	5	5	5	4	3	4

4.

X	4	11	5	7	0	5	9	13	20	6	8	7
Y	7	4	1	11	12	4	2	4	8	9	11	

5.

X	1	5	14	25	80	97	88	65	45	9	1
Y	2	8	13	24	87	95	90	62	41	7	1

6.

X	1	12	15	45	78	102	99	87	64	19	3
Y	5	21	35	56	98	101	78	59	53	17	2

7.

X	2	13	20	24	86	102	109	92	61	18	3
Y	3	12	19	38	85	101	108	88	59	15	2

8.

X	53	87	71	64	78	66	52	54	50	91	55	86	69	82	68
Y	88	84	72	91	89	68	73	52	71	93	87	92	76	72	86

9.

X	3	15	19	25	86	100	110	91	64	20	2
Y	5	11	21	38	85	101	108	88	59	18	1

10.

X	4	18	20	30	95	111	108	88	65	17	1
Y	5	19	22	35	87	112	109	91	61	14	2

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 9
ОДНОФАКТОРНИЙ ДИСПЕРСІЙНИЙ АНАЛІЗ

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для виконання однофакторного дисперсійного аналізу.

Основні поняття: дисперсійний аналіз, рівень значущості, критерій Фішера.

Теоретичні відомості та хід виконання роботи

Дисперсійний аналіз – це статистичний метод обробки результатів вимірювань, які залежать від різних діючих одночасно факторів. Цей метод застосовують для з'ясування питання про суттєвість впливу того чи іншого фактору на вимірювану величину. Залежно від кількості факторів, що вивчаються, розрізняють однофакторний, двофакторний і багатофакторний дисперсійний аналіз.

У дисперсійному аналізі перевірка статистичної значущості відмінності між середніми декількох груп здійснюється на основі вибіркових дисперсій. Ця перевірка здійснюється за допомогою розбиття загальної дисперсії на частини, одна з яких обумовлена випадковою помилкою, а друга – пов'язана з відмінністю середніх значень. Якщо ця відмінність значуща, то нульова гіпотеза щодо існування відмінності між середніми значеннями відкидається на певному рівні значущості.

Методи однофакторного дисперсійною аналізу можна, наприклад, застосувати для перевірки впливу на успішність студентів такого якісного фактору, як організація освітнього процесу для кількох однотипних груп. Кількість рівнів цього фактору дорівнює кількості груп (кількість способів організації).

Сформулюємо задачу однофакторного дисперсійного аналізу для випадку рівної кількості вимірювань на кожному рівні фактору. Нехай досліджується вплив фактору А на певний процес. На кожному рівні А (для кожної варіації параметру) проведено *n* спостережень (див. табл.).

Номери спостережень	**Рівні фактора А**			
	A_1	A_2		A_k
1	x_{11}	x_{21}	…	x_{k1}
2	x_{12}	x_{22}	…	x_{k2}
…	…	…	…	
j	x_{1j}	x_{2j}	…	x_{kj}
…	…	…	…	…
n	x_{1n}	x_{2n}	…	x_{kn}
Σ				

Припускається, що кожна з вибірок (для різних факторів А) має рівні, але невідомі дисперсії та математичні сподівання. У ході аналізу перевіряється гіпотеза про рівність середніх (математичних сподівань), зміст якої полягає в тому, що фактор А не впливає на розподіл випадкових величин X (є випадковим, а не вираженим).

Прийняття рішення щодо гіпотези здійснюється з використанням **критерію Фішера:**

$$\frac{s_A^2}{s_0^2} > F_\alpha[k-1; k(n-1)]$$

де s_A^2 – дисперсія, що характеризує зміну середніх значень для вибірок під впливом фактору А, s_0^2 – дисперсія, що характеризує варіативність поза впливу фактору А, $k - 1$, $k(n-1)$ – ступені вільності F-розподілу.

Розглянемо приклад розрахунків та ключові формули однофакторного дисперсійного аналізу. Потрібно перевірити припущення про те, що фактор швидкості пред'явлення слів впливає на показники їх відтворення.

Досліджуваній ситуації відповідає варіант наступних гіпотез:

H_0: фактор швидкості *не є більш вираженим*, а є випадковим;

H_1: фактор швидкості *є більш вираженим*, ніж випадковим.

Перевірка здійснюється з використанням припущень, що досліджуваний параметр має нормальний розподіл, вибірки незв'язані й однакових обсягів, виміри здійснені за шкалою відношень.

Визначення емпіричного критерію $F_{\text{емп}}$ ґрунтується на співставленні сум за стовпцями із сумою квадратів усіх емпіричних значень. Кожний стовпець – це вибірка, що відповідає певному значенню рівня швидкості.

Введені позначення: $n = 6$ – кількість спостережень (рядків), $k = 3$ – кількість факторів (стовпчиків), $n \cdot k = 18$ – загальна кількість індивідуальних значень, j – індекси рядків, i – індекси стовпчиків.

Розрахункові формули:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2,\ Q_2 = \frac{1}{n}\sum_{i=1}^{k} X_i^2,\ Q_3 = \frac{1}{kn}\left(\sum_{i=1}^{k} X_i\right)^2,$$

де X_i – сума значень у відповідному стовпчику.

$$F_{\text{емп}} = \frac{s_A^2}{s_0^2} = \frac{k(n-1)}{k-1}\frac{Q_2 - Q_3}{Q_1 - Q_2}$$

Нижче наведено виконання розрахунків в MS Excel.

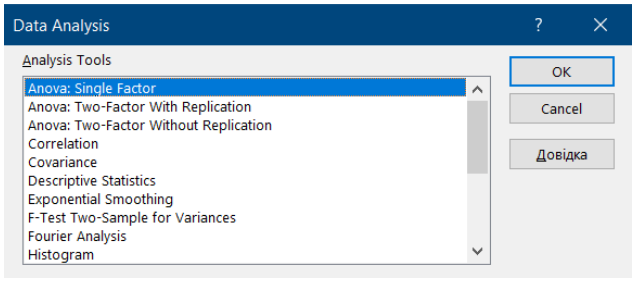
	A	B	C	D	E	F	G
1		**Швидкість пред'явлення**				n	6
2		**Низька**	**Середня**	**Висока**		k	3
3	1	6	5	4			
4	2	7	6	4		Q1	432
5	3	6	5	4		Q2	421
6	4	5	4	3		Q3	410,888889
7	5	6	4	5		$F_{емп}$	6,89393939
8	6	4	5	3			
9	Суми	34	29	23		$F_{0,05}$	3,68232034
10	Середні	5,666667	4,833333	3,833333		$F_{0,01}$	6,35887348

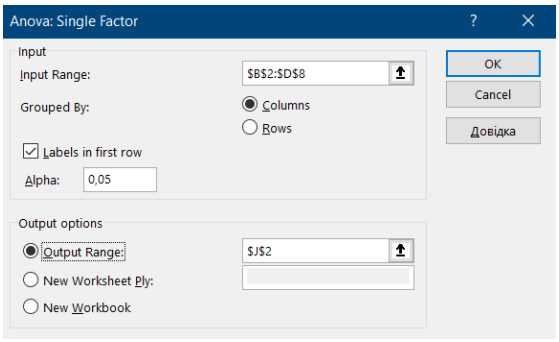
	A	B	C	D	E	F	G
1		**Швидкість пред'явлення**				n	=COUNT(B3:B8)
2		**Низька**	**Середня**	**Висока**		k	=COUNT(B9:D9)
3	1	6	5	4			
4	2	7	6	4		Q1	=SUMSQ(B3:D8)
5	3	6	5	4		Q2	=SUMSQ(B9:D9)/G1
6	4	5	4	3		Q3	=SUM(B9:D9)^2/G1/G2
7	5	6	4	5		$F_{емп}$	=G2*(G1-1)*(G5-G6)/((G2-1)*(G4-G5))
8	6	4	5	3			
9	Суми	=SUM(B3:B8)	=SUM(C3:C8)	=SUM(D3:D8)		$F_{0,05}$	=F.INV(0,95;G2-1;G2*(G1-1))
10	Середні	=AVERAGE(B3:B8)	=AVERAGE(C3:C8)	=AVERAGE(D3:D8)		$F_{0,01}$	=F.INV(0,99;G2-1;G2*(G1-1))

Критичне значення $F_{кр}$ для рівня значущості $\alpha = 0{,}05$ отримується за допомогою функції =F.INV(1-α;k-1;k*(n-1)). Для додаткової перевірки розраховується значення критерію і на рівні значущості $\alpha = 0{,}01$.

Оскільки $F_{емп} > F_{кр}$ (6,89>3,68), нульова гіпотеза відхиляється на рівні значущості 0,05. Таким чином, відмінності в обсязі відтворення слів (фактор швидкості) є більш вираженими (не випадковими).

Розрахунок однофакторної моделі можна здійснити за допомогою пакету Data Analysis (розділ Anova: Single Factor).





Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Низька	6	34	5,666667	1,066667
Середня	6	29	4,833333	0,566667
Висока	6	23	3,833333	0,566667

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	10,11111	2	5,055556	6,893939	0,007527	3,68232
Within Groups	11	15	0,733333			
Total	21,11111	17				

Завдання для самостійного виконання

Використовуючи критерій Фішера на рівні значущості 0,05 методом дисперсійного однофакторного аналізу перевірити нульову гіпотезу про вплив фактора на якість об'єкта на основі п'яти вимірювань для трьох рівнів фактора.

Завдання виконується за варіантами, що відповідають списку групи.

	N	**F1**	**F2**	**F3**		**N**	**F1**	**F2**	**F3**
1	1	14	20	23	**2**	1	7	15	28
	2	18	14	16		2	13	24	9
	3	33	10	15		3	40	43	18
	4	15	6	12		4	12	35	10
	5	10	18	10		5	34	20	35
	N	**F1**	**F2**	**F3**		**N**	**F1**	**F2**	**F3**
3	1	18	10	14	**4**	1	65	30	124
	2	20	8	12		2	64	33	126
	3	30	9	13		3	46	24	123
	4	24	7	18		4	52	38	124
	5	25	6	13		5	53	30	125

	N	F1	F2	F3		N	F1	F2	F3
5	1	31	40	28	**6**	1	7	15	28
	2	33	30	25		2	13	24	9
	3	35	34	32		3	40	43	18
	4	34	34	21		4	12	35	10
	5	34	32	24		5	34	20	35
	N	**F1**	**F2**	**F3**		**N**	**F1**	**F2**	**F3**
7	1	12	8	26	**8**	1	18	14	47
	2	13	10	24		2	40	15	46
	3	11	7	35		3	14	20	48
	4	14	8	22		4	36	21	10
	5	15	9	23		5	35	34	9
	N	**F1**	**F2**	**F3**		**N**	**F1**	**F2**	**F3**
9	1	54	35	14	**10**	1	24	30	9
	2	52	40	32		2	25	41	40
	3	43	24	33		3	48	32	13
	4	44	31	24		4	51	14	15
	5	32	24	12		5	27	25	12

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 10
ЛІНІЙНА КОРЕЛЯЦІЯ

Мета: ознайомитися з розширеними можливостями використання таблчичних процесорів для визначення кореляційних зв'язків.

Основні поняття: випадкові величини, коефіцієнт кореляції, лінійна кореляція.

Теоретичні відомості та хід виконання роботи

Кореляція – це статистична залежність між випадковими величинами, що має імовірнісний характер.

Лінійний кореляційний зв'язок для емпіричних даних, виміряних за шкалою інтервалів або відношень, оцінюється за допомогою коефіцієнта кореляції Пірсона. Він дорівнює сумі добутків відхилень, поділеній на добуток їх стандартних відхилень.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (10.1)$$

де x_i і y_i – значення змінних X і Y; $\bar{x}$ і $\bar{y}$ – середні значення, n – обсяг вибірки.

Формула (1) може бути перетворена, якщо замінити значення змінних x і y нормованими значеннями z_x і z_y:

$$r_{xy} = \frac{\sum_{i=1}^{n}(z_x\cdot z_y)}{n-1} \qquad (10.2)$$

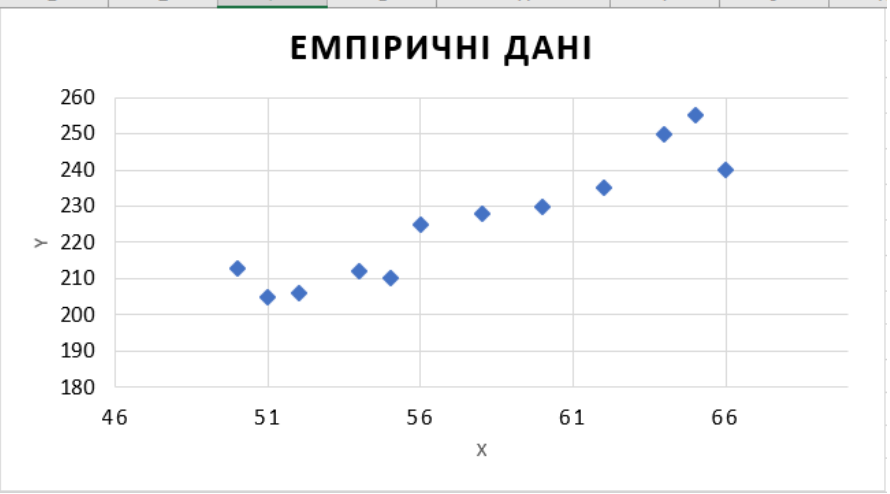
де $z_x = \frac{x-\bar{X}}{s_x}$ і $z_y = \frac{y-\bar{Y}}{s_y}$ – нормовані (на стандартне відхилення) значення змінних X і Y.

Оцінимо зв'язок між змінними X і Y за емпіричними даними, наведеними у таблиці, використавши різні способи, що пропонуються табличними процесорами.

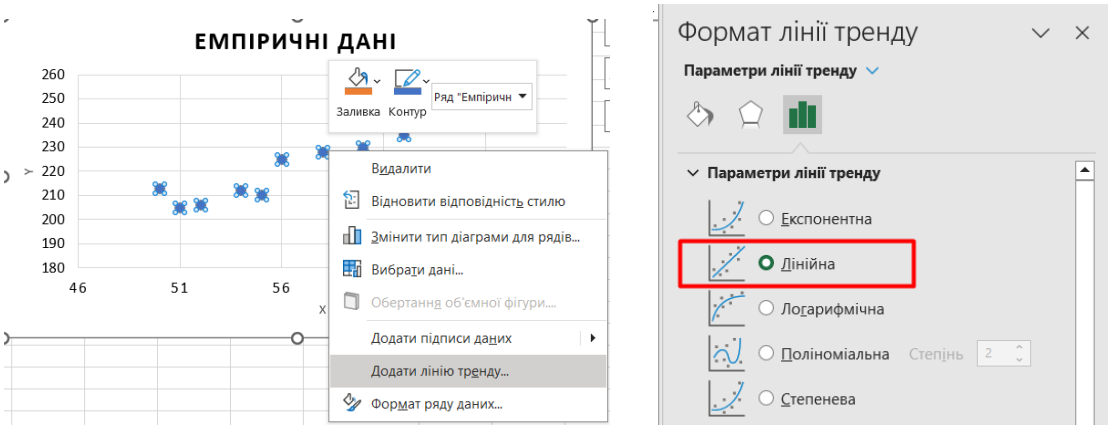
Спосіб 1

1. Створюємо таблицю з емпіричними даними та будуємо графік, обравши тип діаграми Точкова (X Y).

	A	B	C
1	Емпіричні дані		
2		x_i	y_i
3	1	50	213
4	2	52	206
5	3	51	205
6	4	55	210
7	5	56	225
8	6	54	212
9	7	60	230
10	8	58	228
11	9	62	235
12	10	64	250
13	11	65	255
14	12	66	240



Для перевірки лінійності, побудованої залежності, скористаємося опцією Додати лінію тренду (клацнути правою кнопкою мишки по побудованому ряду даних і обрати лінійну лінію тренду).



2. Переконавшись, що кореляція лінійна, розраховуємо коефіцієнт кореляції Пірсона. Для цьог потрібно виконати наступні розрахунки:

- у комірках B16 і C16 розраховуються середні значення;
- у комірках F15 і G15 розраховуються суми квадратів різниць;
- у комірці H18 розраховується сума добутків різниць;
- у комірці B17 розраховується коефіцієнт кореляції

	A	B	C	D	E	F	G	H
1		Емпіричні дані						
2		x_i	y_i	x_i-X_{av}	y_i-Y_{av}	$(x_i$-$X_{av})^2$	$(y_i$-$Y_{av})^2$	$(x_i$-$X_{av})$*$(y_i$-$Y_{av})$
3	1	50	213	=B3-$B$16	=C3-$C$16	=D3^2	=E3^2	=D3*E3
4	2	52	206	=B4-$B$16	=C4-$C$16	=D4^2	=E4^2	=D4*E4
5	3	51	205	=B5-$B$16	=C5-$C$16	=D5^2	=E5^2	=D5*E5
6	4	55	210	=B6-$B$16	=C6-$C$16	=D6^2	=E6^2	=D6*E6
7	5	56	225	=B7-$B$16	=C7-$C$16	=D7^2	=E7^2	=D7*E7
8	6	54	212	=B8-$B$16	=C8-$C$16	=D8^2	=E8^2	=D8*E8
9	7	60	230	=B9-$B$16	=C9-$C$16	=D9^2	=E9^2	=D9*E9
10	8	58	228	=B10-$B$16	=C10-$C$16	=D10^2	=E10^2	=D10*E10
11	9	62	235	=B11-$B$16	=C11-$C$16	=D11^2	=E11^2	=D11*E11
12	10	64	250	=B12-$B$16	=C12-$C$16	=D12^2	=E12^2	=D12*E12
13	11	65	255	=B13-$B$16	=C13-$C$16	=D13^2	=E13^2	=D13*E13
14	12	66	240	=B14-$B$16	=C14-$C$16	=D14^2	=E14^2	=D14*E14
15	Суми	=SUM(B3:B14)	=SUM(C3:C14)	=SUM(D3:D14)	=SUM(E3:E14)	=SUM(F3:F14)	=SUM(G3:G14)	=SUM(H3:H14)
16	Середні	=AVERAGE(B3:B14)	=AVERAGE(C3:C14)					
17	rxy	=H15/SQRT(F15*G15)						

15	Суми	693	2709	0	0	346	3176	979,25
16	Середні	58	226					
17	rxy	0,9337731						

Отримане значення коефіцієнта кореляції +0.93 свідчить про суттєвий прямий зв'язок між ознаками.

Спосіб 2

Розрахункові формули та отримані результати наведено на рисунках нижче. Розраховуємо наступні величини:

- у комірках B16 і C16 розраховуються середні значення;
- у комірках B17 і C17 розраховуються стандартні відхилення;
- у стовпчиках D і E розраховуються нормовані дані (зверніть увагу, що середнє для нормованих даних дорівнює 0, а стандартне відхилення – 1).

- У комірці B18 розрахувати коефіцієнт кореляції за формулою 2.

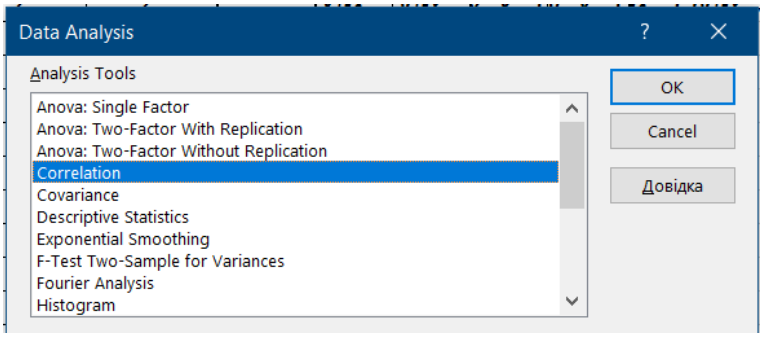
	A	B	C	D	E
1		**Емпіричні дані**			
2		x_i	y_i	z_x	z_y
3	1	50	213	=(B3-$B$16)/$B$17	=(C3-$C$16)/$C$17
4	2	52	206	=(B4-$B$16)/$B$17	=(C4-$C$16)/$C$17
5	3	51	205	=(B5-$B$16)/$B$17	=(C5-$C$16)/$C$17
6	4	55	210	=(B6-$B$16)/$B$17	=(C6-$C$16)/$C$17
7	5	56	225	=(B7-$B$16)/$B$17	=(C7-$C$16)/$C$17
8	6	54	212	=(B8-$B$16)/$B$17	=(C8-$C$16)/$C$17
9	7	60	230	=(B9-$B$16)/$B$17	=(C9-$C$16)/$C$17
10	8	58	228	=(B10-$B$16)/$B$17	=(C10-$C$16)/$C$17
11	9	62	235	=(B11-$B$16)/$B$17	=(C11-$C$16)/$C$17
12	10	64	250	=(B12-$B$16)/$B$17	=(C12-$C$16)/$C$17
13	11	65	255	=(B13-$B$16)/$B$17	=(C13-$C$16)/$C$17
14	12	66	240	=(B14-$B$16)/$B$17	=(C14-$C$16)/$C$17
15	Суми	=SUM(B3:B14)	=SUM(C3:C14)		
16	Середні	=AVERAGE(B3:B14)	=AVERAGE(C3:C14)		
17	Ст.відх.	=STDEV.S(B3:B14)	=STDEV.S(C3:C14)	=STDEV.S(D3:D14)	=STDEV.S(E3:E14)
18	rxy	=SUMPRODUCT(D3:D14;E3:E14)/(COUNT(A3:A14)-1)			

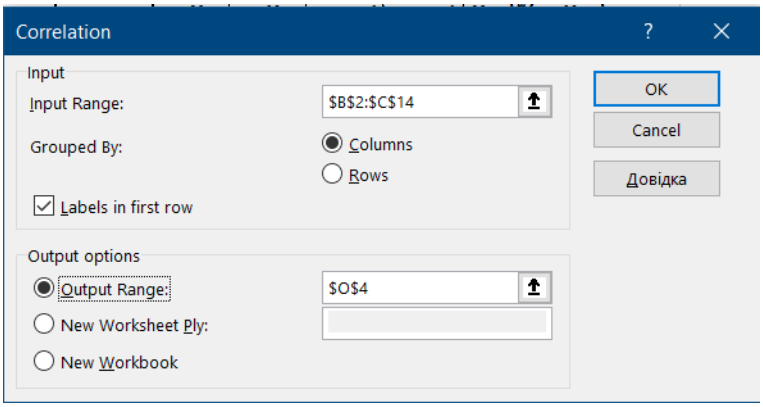
	A	B	C	D	E
1		**Емпіричні дані**		**Нормовані дані**	
2		x_i	y_i	z_x	z_y
3	1	50	213	-1	-1
4	2	52	206	-1	-1
5	3	51	205	-1	-1
6	4	55	210	0	-1
7	5	56	225	0	0
8	6	54	212	-1	-1
9	7	60	230	0	0
10	8	58	228	0	0
11	9	62	235	1	1
12	10	64	250	1	1
13	11	65	255	1	2
14	12	66	240	1	1
15	Суми	693	2709	0	0
16	Середні	58	226	0	0
17	Ст.відх.	6	17	1	1
18	rxy	0,933773082			

Спосіб 3

Найшвидшим способом розрахунку коефіцієнта кореляції Пірсона є використання вбудованої функції =PEARSON(B3:B14;C3:C14).

Обрахувати значення коефіцієнта кореляції можна також з використанням розширення Data Analysis.





	xi	*yi*
xi	1	
yi	0,933773	1

За умови невеликої кількості пар ознак X та Y (n<30) у практичній діяльності використовують таку інтерпретацію значення коефіцієнта кореляції (інтерпретація американського вченого Чеддока):

- якщо 0,90≤r≤0,99, між ознаками існує дуже високий ступінь зв'язку;
- якщо 0,7≤r≤0,9, то між ознаками існує високий ступінь зв'язку;
- якщо 0,50≤r≤0,69, то між ознаками існує помірний ступінь зв'язку;
- якщо 0,2≤r≤0,49, між ознаками існує слабкий ступінь взаємозв'язку.

Коефіцієнт детермінації – це числове значення, що отримують в результаті піднесення до другого степеня значення коефіцієнта кореляції. Значення коефіцієнта детермінації вказує на вплив загальних факторів на досліджувані ознаки.

Коефіцієнт детермінації завжди позитивний і перебуває в межах від нуля до одиниці. Він показує долю варіації результативної ознаки Y під впливом факторної ознаки X.

Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи.

Побудувати діаграму розсіяння, розрахувати усіма описаними в роботі способами коефіцієнт кореляції Пірсона та коефіцієнт детермінації, оцінити рівень значущості зв'язку між ознаками X та Y.

	Завдання													
1	X	2,06	2,58	3,14	3,54	4,18	4,78	5,11	5,67	6,02	6,65	7,05	7,52	8,03	8,56
	Y	14,87	15,78	16,79	18,03	18,29	19,93	20,32	21,18	2,47	23,47	24,07	25,57	27,07	27,62
2	X	2,53	3,54	3,84	3,84	4,22	4,81	6,53	5,82	6,43	7,73	8,19	7,65	9,31	9,26
	Y	19,66	20,53	21,31	22,59	23,27	24,44	25,85	26,74	27,36	28,37	29,22	30,5	31,21	32,56
3	X	2,17	2,9	3,29	4,13	5,25	4,92	5,79	5,87	6,99	7,04	8,14	8,06	8,57	9,45
	Y	12,5	13,88	15,16	16,06	16,66	17,65	18,46	19,54	20,58	21,77	22,15	23,8	24,79	25,57
4	X	3,65	3,82	3,76	5,24	5,03	5,52	5,62	6,98	6,91	7,95	7,24	9,27	8,46	10,3
	Y	10,22	10,58	12,01	12,84	13,28	15,13	15,84	17,08	17,99	18,32	19,49	20,59	21,35	23,2
5	X	3,22	3,87	4,95	5,1	5,98	7,28	6,9	7,54	7,91	8,4	8,14	8,76	9,67	10,28
	Y	16,62	17,63	19,22	19,36	20,52	21,95	22,45	23,56	24,9	25,53	26,11	28,02	28,37	29,48

6	X	2,16	2,65	3,49	3,16	3,85	4,58	5,33	5,89	6,2	6,39	6,95	7,25	7,8	8,47
	Y	15,21	15,42	16,44	17,93	18,52	19,8	20,76	21,3	22,25	24,14	24,17	25,66	26,5	27,46
7	X	4,57	5,42	5,29	6,33	7,63	7,53	7,73	8,44	9,49	9,18	10,14	9,94	10,92	11,89
	Y	12,11	12,3	13,82	14,84	15,86	16,41	17,8	18,61	19,57	21,26	21,08	22,99	23,43	24,63
8	X	2,25	2,98	2,15	2,71	3,07	4,59	4,77	5,34	5,45	6,0	6,25	6,79	8,24	8,51
	Y	16,21	17,75	16,39	18,87	19,6	21,21	21,84	23,0	24,44	25,36	25,54	27,14	27,95	28,99
9	X	6,15	5,66	7,5	6,9	8,31	8,25	9,39	9,73	9,33	10,5	11,1	11,51	12,42	12,4
	Y	10,89	11,92	12,45	13,27	14,12	15,23	16,07	17,4	18,68	19,46	20,52	21,32	22,58	23,73
10	X	1,86	1,91	2,14	3,39	3,95	4,3	5,1	5,47	5,97	6,16	6,46	6,07	6,71	7,16
	Y	7,24	8,02	9,28	10,12	11,12	12,19	13,01	14,12	15,21	16,29	17,01	18,03	19,19	20,21

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 11
НЕЛІНІЙНА КОРЕЛЯЦІЯ

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для визначення нелінійних кореляційних зв’язків.

Основні поняття: випадкові величини, коефіцієнт кореляції, нелінійна кореляція.

Теоретичні відомості та хід виконання роботи

Кореляція – це статистична залежність між випадковими величинами, що має імовірнісний характер. Нелінійна кореляція вказує на взаємозв'язок між двома змінними, який не може бути описаний простою лінійною функцією. У випадку нелінійної кореляції, зміна в одній змінній не призводить до постійної зміни в іншій змінній, а зміни можуть бути більш складними та непередбачуваними.

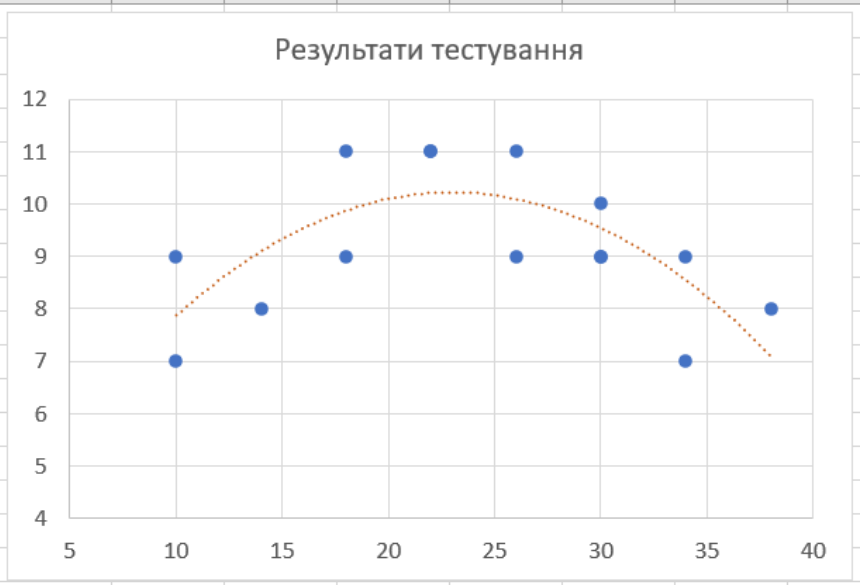
Уявімо, що ми досліджуємо взаємозв'язок між кількістю годин витрачених на вивчення та результатами іспиту. На перший погляд, можна припустити, що більше годин витрачених на вивчення призведе до кращих результатів на іспиті. Проте, якщо ви врахуєте, що перевищення певної кількості годин може викликати стрес або втому, то зв'язок між кількістю годин і результатами іспиту може бути нелінійним.

Наприклад, для перших кількох годин вивчення може спостерігатися значний зростання результатів на іспиті, але після деякої точки додаткові години вивчення можуть не приносити значного покращення результатів, а можуть навіть призвести до зниження результатів через втому чи стрес.

Таким чином, в цьому прикладі між кількістю годин вивчення та результатами іспиту існує нелінійна кореляція, яка не може бути адекватно описана простою лінійною функцією.

Розглянемо послідовність розрахунку коефіцієнта нелінійної кореляції на прикладі, що наводиться нижче (Руденко, 2012). Потрібно оцінити зв’язок між віком (змінна X) і результатами тесту «цифра-знак» шкали інтеоекту дорослих Векслера (змінна Y). Упорядковані за віком дані 15 осіб наведено в таблиці.

	A	B	C
1		**Вік**	**Тест**
2	***i***	x_i	y_i
3	1	10	7
4	2	10	9
5	3	14	8
6	4	18	9
7	5	18	11
8	6	22	11
9	7	22	11
10	8	26	9
11	9	26	11
12	10	30	9
13	11	30	9
14	12	30	10
15	13	34	7
16	14	34	9
17	15	38	8



1. Оцінюємо характер лінійності (нелінійності) зв'язку між значеннями ознак віку (X) і тесту (Y) за допомогою діаграми розсіяння.

2. Переконуємося, що кореляція нелінійна – спочатку результати тестування круто зростають для осіб віком від 10 до 22 років, досягають максимального значення, а потім повільно зменшуються.

Якісна картина дає підстави для застосування кількісної міри нелінійності – кореляційного відношення, чисельне значення якого знаходиться в межах від 0 до 1. Розрахунок здійснюється за формулою:

$$\eta^2_{y,x} = 1 - \frac{SS_{\text{внутр}}}{SS_{\text{загал}}}$$

де $SS_{\text{внутр}} = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} (y_i - \bar{y})^2$ – внутрішньогрупова сума квадратів відхилень y_i від середнього; $SS_{\text{загал}} = s_y^2 \cdot (n-1)$ – загальна сума квадратів.

3. Розраховуємо квадрати різниць s_i окремо для кожної вікової групи (вікові групи виділено зафарбованими рядками). Перша вікова група – 10 років (включає два значення), друга – 14 (одне значення), третя – 18 (два значення), четверта – 22 (два значення), п'ята – 26 (два значення), шоста – 30 (три значення), сьома – 34 (два значення), восьма – 38 (одне значення).

4. У комірці D18 розрахувати $SS_{\text{внутр}}$ (=СУММ(D3:D17)).

5. У комірці D19 розрахувати $SS_{\text{загал}}$ (=ДИСП(C3:C17)*(A17-1)).

6. У комірці D20 отримати відношення $\eta^2_{y,x}$ (=1-D18/D19).

7. У комірці D21 розрахувати коефіцієнт кореляції Пірсона для всього масиву за допомогою функції MS Excel =PEARSON(B3:B17;C3:C17). Отримане значення дорівнюватиме приблизно нулю (-0,04), що свідчить про (нібито) відсутність зв'язку між змінними.

8. Розрахувати коефіцієнти кореляції окремо для частин масиву: у комірці D22 для віку від 10 до 22, у комірці D23 для віку від 26 до 38.

Формули для розрахунків і отримані результати наведено на рисунку нижче.

	A	B	C	D
1		**Вік**	**Тест**	
2	***i***	x_i	y_i	s_i
3	1	10	7	=(C3-AVERAGE($C$3:$C$4))^2
4	2	10	9	=(C4-AVERAGE($C$3:$C$4))^2
5	3	14	8	=(C5-AVERAGE($C$5:$C$5))^2
6	4	18	9	=(C6-AVERAGE($C$6:$C$7))^2
7	5	18	11	=(C7-AVERAGE($C$6:$C$7))^2
8	6	22	11	=(C8-AVERAGE($C$8:$C$9))^2
9	7	22	11	=(C9-AVERAGE($C$8:$C$9))^2
10	8	26	9	=(C10-AVERAGE($C$10:$C$11))^2
11	9	26	11	=(C11-AVERAGE($C$10:$C$11))^2
12	10	30	9	=(C12-AVERAGE($C$12:$C$14))^2
13	11	30	9	=(C13-AVERAGE($C$12:$C$14))^2
14	12	30	10	=(C14-AVERAGE($C$12:$C$14))^2
15	13	34	7	=(C15-AVERAGE($C$15:$C$16))^2
16	14	34	9	=(C16-AVERAGE($C$15:$C$16))^2
17	15	38	8	=(C17-AVERAGE($C$17:$C$17))^2
18	$SS_{\text{внутр}}$			=SUM(D3:D17)
19	$SS_{\text{загал}}$			=VAR.S(C3:C17)*(A17-1)
20	Кореляційне відношення η^2_{xy}			=1-D18/D19
21	Коефіцієнти кореляції для віку		від 10 до 18	=PEARSON(B3:B17;C3:C17)
22			від 10 до 22	=PEARSON(B3:B9;C3:C9)
23			від 26 до 38	=PEARSON(B10:B17;C10:C17)

	A	B	C	D
1		**Вік**	**Тест**	
2	***i***	x_i	y_i	s_i
3	1	10	7	1
4	2	10	9	1
5	3	14	8	0
6	4	18	9	1
7	5	18	11	1
8	6	22	11	0
9	7	22	11	0
10	8	26	9	1
11	9	26	11	1
12	10	30	9	0,11
13	11	30	9	0,11
14	12	30	10	0,44
15	13	34	7	1
16	14	34	9	1
17	15	38	8	0
18	$SS_{\text{внутр}}$			8,67
19	$SS_{\text{загал}}$			26,40
20	Кореляційне відношення η^2_{xy}			0,67
21	Коефіцієнти кореляції для віку		від 10 до 18	-0,04
22			від 10 до 22	0,83
23			від 26 до 38	-0,69

Отримане для віку від 10 до 22 років значення коефіцієнта кореляції має високе додатне значення (+0,83), що підтверджує прямий зв’язок, який можна спостерігати на діаграмі. Для віку від 26 до 38 років коефіцієнт кореляції має від’ємне значення (-0,69), що інтерпретується як зворотний зв’язок. Значення кореляційного відношення $\eta_{y,x}^2 = 0{,}67$ підтверджує високий рівень нелінійності зв'язку змінних X і Y.

Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи. Побудувати діаграму розсіяння, розрахувати коефіцієнт кореляції, оцінити рівень значущості зв’язку між масивами даних.

Варіант 1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
xi	10	10	10	10	10	14	14	14	14	18	18	18	18	22
yi	7	8	9	9	10	8	9	10	11	9	10	11	12	11
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
xi	22	22	22	26	26	26	30	30	30	30	34	34	34	38
yi	11	12	12	9	10	11	8	9	9	10	7	9	10	8

Варіант 2

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
xi	12	12	12	12	12	16	16	16	16	20	20	20	20	24
yi	9	10	11	11	12	10	11	12	13	11	12	13	14	13
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
xi	24	24	24	28	28	28	32	32	32	32	36	36	36	40
yi	13	14	14	11	12	13	10	11	11	12	9	11	12	10

Варіант 3

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
xi	9	9	9	9	9	13	13	13	13	17	17	17	17	21
yi	6	7	8	8	9	7	8	9	10	8	9	10	11	10
i	12	13	14	15	16	17	18	19	20	21	22	23	24	25
xi	21	21	21	25	25	25	29	29	29	29	33	33	33	37
yi	10	11	11	8	9	10	7	8	8	9	6	8	9	7

Варіант 4

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
xi	14	14	14	14	14	18	18	18	18	22	22	22	22	26
yi	11	12	13	13	14	12	13	14	15	13	14	15	16	15
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
xi	26	26	26	30	30	30	34	34	34	34	38	38	38	42
yi	15	16	16	13	14	15	12	13	13	14	11	13	14	12

Варіант 5

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
xi	13	13	13	13	13	17	17	17	17	21	21	21	21	25
yi	10	11	12	12	13	11	12	13	14	12	13	14	15	14
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
xi	25	25	25	29	29	29	33	33	33	33	37	37	37	41
yi	14	15	15	12	13	14	11	12	12	13	10	12	13	11

6

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	10	10	10	10	10	14	14	14	14	18	18	18	18	22
y_i	7	8	9	9	10	8	9	10	11	9	10	11	12	11
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
x_i	22	22	22	26	26	26	30	30	30	30	34	34	34	38
y_i	11	12	12	9	10	11	8	9	9	10	7	9	10	8

7

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	12	12	12	12	12	16	16	16	16	20	20	20	20	24
y_i	9	10	11	11	12	10	11	12	13	11	12	13	14	13
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
x_i	24	24	24	28	28	28	32	32	32	32	36	36	36	40
y_i	13	14	14	11	12	13	10	11	11	12	9	11	12	10

8

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	9	9	9	9	9	13	13	13	13	17	17	17	17	21
y_i	6	7	8	8	9	7	8	9	10	8	9	10	11	10
i	12	13	14	15	16	17	18	19	20	21	22	23	24	25
x_i	21	21	21	25	25	25	29	29	29	29	33	33	33	37
y_i	10	11	11	8	9	10	7	8	8	9	6	8	9	7

9

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	14	14	14	14	14	18	18	18	18	22	22	22	22	26
y_i	11	12	13	13	14	12	13	14	15	13	14	15	16	15
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
x_i	26	26	26	30	30	30	34	34	34	34	38	38	38	42
y_i	15	16	16	13	14	15	12	13	13	14	11	13	14	12

10

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	13	13	13	13	13	17	17	17	17	21	21	21	21	25
y_i	10	11	12	12	13	11	12	13	14	12	13	14	15	14
i	15	16	17	18	19	20	21	22	23	24	25	26	27	28
x_i	25	25	25	29	29	29	33	33	33	33	37	37	37	41
y_i	14	15	15	12	13	14	11	12	12	13	10	12	13	11

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 12
ОЦІНКА МІР ВЗАЄМОЗВ'ЯЗКУ ОЗНАК

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для визначення кореляційних зв'язків багатьох груп ознак.

Основні поняття: випадкові величини, коефіцієнт взаємної зв'язаності Чупрова, коефіцієнт взаємної зв'язаності Пірсона.

Теоретичні відомості та хід виконання роботи

Коефіцієнти взаємної зв'язаності використовуються для оцінки зв'язку в ситуаціях, коли кожна якісна ознака складається більш ніж з двох груп.

Припустимо, ви досліджуєте зв'язок між рівнем освіти та заробітною платою. Обидві ці змінні є категоріальними, і в кожній з них є більше двох категорій. Ви зібрали дані від 200 респондентів і поділили їх за рівнем освіти (наприклад, вища, середня, початкова) та за категоріями заробітної платні (наприклад, висока, середня, низька).

Оцінка мір взаємозв'зку ознак використовується в багатьох областях, таких як економіка, соціологія, психологія та інші, для вивчення взаємозв'язку між різними змінними.

Для оцінки зв'язку між цими двома ознаками можна обчислити коефіцієнти кореляції Чупрова (Q-коефіцієнт) або Пірсона.

Коефіцієнт Чупрова К використовується у випадку неоднакової кількості рядків і стовпчиків таблиці спряженості ($k_1 \neq k_2$):

$$K = \sqrt{\frac{\varphi^2}{\sqrt{(k_1 - 1) \cdot (k_2 - 1)}}}$$

де k_1 і k_2 – кількість груп першої і другої ознаки (параметри X і Y).

Коефіцієнт Чупрова варіюється від 0 до 1, де значення ближче до 1 вказує на сильніший зв'язок між категоріальними змінними, тоді як значення ближче до 0 показує слабший зв'язок. Цей коефіцієнт допомагає визначити, наскільки сильно змінна одного типу пов'язана з категоріями іншого типу.

Коефіцієнт Пірсона С використовується, коли кількість рядків і кількість стовпчиків у таблиці спряженості збігаються ($k_1 = k_2$):

$$\mathrm{C} = \sqrt{\frac{\varphi^2}{1 + \varphi^2}}$$

де

$$\varphi^2 = \sum_{y=1}^{k1} \left(\frac{\sum_{x=1}^{k2} \left(\frac{n_{xy}^2}{n_x} \right)}{n_y} \right) - 1$$

і n_{xy} – значення у таблиці на перетині рядка x і стовпчика y, n_x – сума значень n_{xy} у певному стовпчику, n_y – сума значень n_{xy} у певному рядку.

У наведеному нижче прикладі потрібно оцінити зв'язаність між приналежністю осіб до певної соціальної групи та їх станами, за даними, які представлено у таблиці.

Соціальні групи (параметр Y)	Можливі стани (параметр X)				Всього
	Стан 1	Стан 2	Стан 3	Стан 4	
Студенти	3	9	3	1	16
Службовці	8	2	4	2	16
Пенсіонери	3	1	1	8	13
Всього	14	12	8	11	45

Послідовність рішення:

1. Для ситуацією з різною кількістю рядків і стовпчиків використовуємо коефіцієнт взаємної зв'язаності Чупрова К. У нашому випадку, кількість рядків $k_1 = 3$, а кількість стовпчиків $k_2 = 4$.

2. Записуємо вхідні дані в таблицю, розраховуємо суми значень по рядках і по стовпчиках.

3. Розписуємо вираз для φ^2, виходячи з умови $k_1 = 3$, $k_2 = 4$:

$$\varphi^2 = \sum_{y=1}^{k1}\left(\frac{\sum_{x=1}^{k2}\left(\frac{n_{xy}^2}{n_x}\right)}{n_y}\right) - 1 = \frac{\sum_{x=1}^{4}\left(\frac{n_{xy}^2}{n_x}\right)}{n_{y1}} + \frac{\sum_{x=1}^{4}\left(\frac{n_{xy}^2}{n_x}\right)}{n_{y2}} + \frac{\sum_{x=1}^{4}\left(\frac{n_{xy}^2}{n_x}\right)}{n_{y3}} - 1$$

4. Розраховуємо окремі складові φ^2 і знайдемо їх суму: $\varphi^2 = A_1 + A_2 + A_3 = 0{,}505$.

$$A_1 = \frac{\sum_{x=1}^{4}\left(\frac{n_{xy}^2}{n_x}\right)}{n_{y1}} = \frac{\frac{3^2}{14} + \frac{9^2}{12} + \frac{3^2}{8} + \frac{1^2}{11}}{16} = 0{,}538$$

$$A_2 = \frac{\sum_{x=1}^{4}\left(\frac{n_{xy}^2}{n_x}\right)}{n_{y1}} = \frac{\frac{8^2}{14} + \frac{2^2}{12} + \frac{4}{8} + \frac{2^2}{11}}{16} = 0{,}454$$

$$A_3 = \frac{\sum_{x=1}^{4}\left(\frac{n_{xy}^2}{n_x}\right)}{n_{y1}} = \frac{\frac{3^2}{14} + \frac{1^2}{12} + \frac{1^2}{8} + \frac{8^2}{11}}{16} = 0{,}513$$

5. Визначити параметр коефіцієнта зв'язаності К.

Формули для розрахунків та отримані результати наведені на рисунку.

Соціальні групи (параметр Y)	Можливі стани (параметр X)				Всього
	Стан 1	Стан 2	Стан 3	Стан 4	
Студенти	3	9	3	1	=SUM(C4:F4)
Службовці	8	2	4	2	=SUM(C5:F5)
Пенсіонери	3	1	1	8	=SUM(C6:F6)
Всього	=SUM(C4:C6)	=SUM(D4:D6)	=SUM(E4:E6)	=SUM(F4:F6)	=SUM(G4:G6)
k1	=COUNT(C4:C6)				
k2	=COUNT(C4:F4)				
A1	=(C4^2/C$7+D4^2/D$7+E4^2/E$7+F4^2/F$7)/G4				
A2	=(C5^2/C$7+D5^2/D$7+E5^2/E$7+F5^2/F$7)/G5				
A3	=(C6^2/C$7+D6^2/D$7+E6^2/E$7+F6^2/F$7)/G6				
ϕ2	=SUM(C10:C12)-1				
K	=SQRT((C13)/SQRT((C8-1)*(C9-1)))				

Соціальні групи (параметр Y)	Можливі стани (параметр X)				Всього
	Стан 1	Стан 2	Стан 3	Стан 4	
Студенти	3	9	3	1	16
Службовці	8	2	4	2	16
Пенсіонери	3	1	1	8	13
Всього	14	12	8	11	45
k1	3				
k2	4				
A1	0,538048				
A2	0,454275				
A3	0,513029				
$\phi 2$	0,505351				
K	0,454212				

Значення коефіцієнта взаємної зв’язаності Чупрова К=0,45 свідчить про помірну взаємну зв’язаність між параметрами X і Y. Напрямок зв’язаності коефіцієнт К не вказує. Це можна оцінити за формою спільного розподілу.

Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи. Розрахувати коефіцієнт взаємної зв’язаності для заданих даних та зробити високвок про силу зв'зку.

Варіант 1

Параметри Y	Параметри X			Всього
	X1	X2	X3	
Y1	8	4	6	
Y2	12	6	4	
Y3	11	3	1	
Y4	15	6	1	
Всього				

Варіант 2

Параметри Y	Параметри X				Всього
	X1	X2	X3	X4	
Y1	10	43	7	5	
Y2	42	14	25	6	
Y3	51	11	10	33	
Всього					

Варіант 3

Параметри Y	Параметри X			Всього
	X1	X2	X3	
Y1	74	12	5	
Y2	71	11	12	
Y3	62	21	23	
Всього				

Варіант 4

Параметри Y	Параметри X			Всього
	X1	X2	X3	
Y1	8	1	6	
Y2	88	259	154	
Y3	13	15	4	
Y4	14	3	3	
Всього				

5

Параметри Y	Параметри X			Всього
	X1	X2	X3	
Y1	2	2	5	
Y2	9	8	88	
Y3	33	34	37	
Всього				

6

Параметри Y	Параметри X				Всього
	X1	X2	X3	X4	
Y1	10	43	7	5	
Y2	42	14	25	6	
Y3	51	11	10	33	
Всього					

7

Параметри Y	Параметри X			Всього
	X1	X2	X3	
Y1	8	1	6	
Y2	88	259	154	
Y3	13	15	4	
Y4	14	3	3	
Всього				

8

Параметри Y	Параметри X			Всього
	X1	X2	X3	
Y1	2	2	5	
Y2	9	8	88	
Y3	33	34	37	
Всього				

9

Параметри Y	Параметри X			Всього
	X1	X2	X3	
Y1	74	12	5	
Y2	71	11	12	
Y3	62	21	23	
Всього				

10

Параметри Y	Параметри X			Всього
	X1	X2	X3	
Y1	8	4	6	
Y2	12	6	4	
Y3	11	3	1	
Y4	15	6	1	
Всього				

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 13
ОДНОВИМІРНА ЛІНІЙНА РЕГРЕСІЯ

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для визначення коефіцієнта одновимірної лінійної регресії.

Основні поняття: лінійна регресія, одновимірна регресія.

Теоретичні відомості та хід виконання роботи

Одновимірна лінійна регресія – це метод аналізу, що використовується для вивчення взаємозв'язку між двома змінними, де одна змінна (незалежна змінна) використовується для прогнозування значень іншої змінної (залежна змінна), при цьому вважається, що взаємозв'язок між ними може бути виражений лінійною функцією. У випадку одновимірної лінійної регресії ми маємо тільки одну незалежну змінну, яка використовується для прогнозування значень залежної змінної.

Побудова лінійної регресії полягає у розрахунках коефіцієнтів лінійної регресії a_0 і a_1:

$$a_1 = \frac{\sum(x_i - \bar{X}) \cdot (y_i - \bar{Y})}{\sum(x_i - \bar{X})^2}$$

$$a_0 = \bar{Y} - a_1 \cdot \bar{X}$$

де $\bar{X}$ і $\bar{Y}$ – середні значення змінних X і Y.

Вибір значень коефіцієнтів a_0 і a_1 виконується за методом найменших квадратів так, щоб сума $\sum(y_i - \bar{Y})^2 = \sum(y_i - a_0 - a_1 \cdot \bar{X})^2$ була мінімальною.

Приклад використання одновимірної лінійної регресії може включати прогнозування вартості нерухомості (залежна змінна) на основі площі будинку (незалежна змінна). У цьому випадку площа будинку буде використовуватися для прогнозування ціни будинку. За допомогою методу лінійної регресії ми можемо побудувати лінію, яка найкращим чином відповідає спостереженим даним, і застосувати цю лінію для прогнозування вартості будинку на основі його площі.

У прикладі, наведеному нижче, потрібно оцінити залежність успішності навчання (Y) від затраченого часу (X), за даними, які представлено у таблиці.

	A	B	C
1	**Емпіричні дані**		
2		x_i	y_i
3	1	2	5
4	2	3	5
5	3	2	3
6	4	3	6
7	5	3	3
8	6	1	3
9	7	3	5
10	8	2	3
11	9	3	5
12	10	3	5
13	11	2	2

Послідовність рішення з розрахунку коефіцієнтів регресії a_0 і a_1:

1. У комірки B15 і C15 ввести =AVERAGE(B3:B13) і = AVERAGE(C3:C13) й отримати середні значення масивів $\bar{X} = 2{,}39$, $\bar{Y} = 4{,}09$.
2. У комірках D13:H13 розрахувати різниці, добутки і квадрати різниць за допомогою відповідних формул.
3. У комірках F14:H14 розрахувати суми добутків і квадратів різниць.
4. У комірках D17 і D18 розрахувати коефіцієнти лінійної регресії a_0 і a_1 за допомогою виразів =F14/G14 і =C15-D17*B15: $a_0 = 1{,}37$, $a_1 = 0{,}82$.
5. Виконати у комірках I3:I13 розрахунки теоретичного значення $\tilde{Y} = a_0 + a_1 \cdot X$ за рівнянням $\tilde{Y} = 0{,}82 + 1{,}37 \cdot X$. Для цього в комірку I3 ввести вираз =$D$18+$D$17*B3. Аналогічні вирази ввести в інші комірки стовпчика I.

Формули для розрахунків та отримані результати наведені на рисунках.

	A	B	C	D	E	F	G	H	I	J
1	**Емпіричні дані**			**Розрахунки**					**Регресія**	
2		x_i	y_i	x_i-X_{av}	y_i-Y_{av}	(x_i-X_{av})*(y_i-Y_{av})	(x_i-X_{av})^2	(y_i-Y_{av})^2	*Y*	*X*
3	1	2,10	5,00	-0,29	0,91	-0,26	0,08	0,83	3,69	2,77
4	2	3,40	5,00	1,01	0,91	0,92	1,02	0,83	5,47	2,77
5	3	1,50	3,00	-0,89	-1,09	0,97	0,79	1,19	2,87	1,93
6	4	2,90	6,00	0,51	1,91	0,97	0,26	3,64	4,79	3,19
7	5	2,50	3,00	0,11	-1,09	-0,12	0,01	1,19	4,24	1,93
8	6	1,40	3,00	-0,99	-1,09	1,08	0,98	1,19	2,73	1,93
9	7	2,50	5,00	0,11	0,91	0,10	0,01	0,83	4,24	2,77
10	8	2,20	3,00	-0,19	-1,09	0,21	0,04	1,19	3,83	1,93
11	9	3,00	5,00	0,61	0,91	0,55	0,37	0,83	4,93	2,77
12	10	3,30	5,00	0,91	0,91	0,83	0,83	0,83	5,34	2,77
13	11	1,50	2,00	-0,89	-2,09	1,86	0,79	4,37	2,87	1,51
14	**Суми**					7,11	5,19	16,91		
15	**Середні**	2,39	4,09							
16		**r_{xy}**		0,76						
17		**a_1**		1,37		**b_1**	0,42			
18		**a_0**		0,82		**b_0**	0,67			

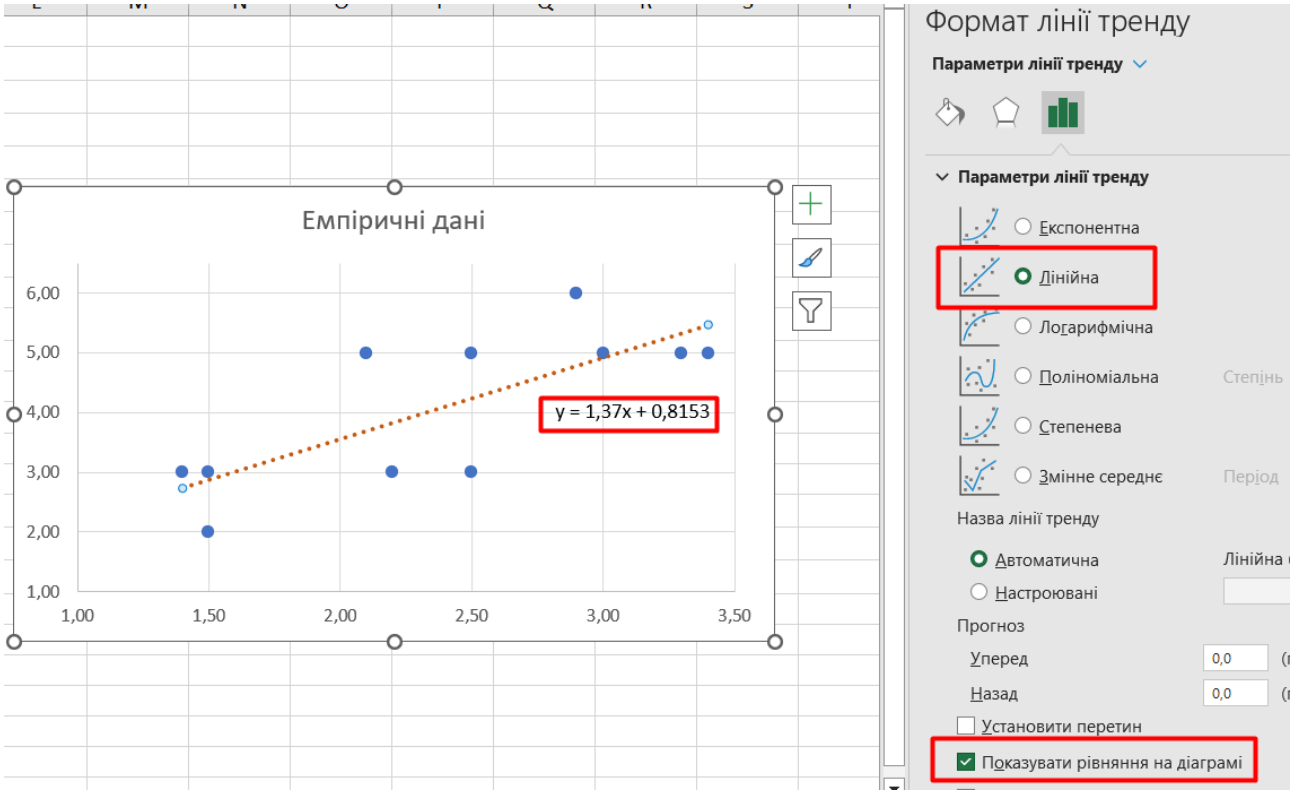
A	B	C	D	E	F	G	H	I	J
Емпіричні дані			**Розрахунки**					**Регресія**	
	x_i	y_i	x_i-X_{av}	y_i-Y_{av}	(x_i-X_{av})*(y_i-Y_{av}	(x_i-X_{av})^2	(y_i-Y_{av})^2	*Y*	*X*
1	2,1	5	=B3-$B$15	=C3-$C$15	=D3*E3	=D3^2	=E3^2	=$D$18+$D$17*B3	=$G$18+$G$17*C3
2	3,4	5	=B4-$B$15	=C4-$C$15	=D4*E4	=D4^2	=E4^2	=$D$18+$D$17*B4	=$G$18+$G$17*C4
3	1,5	3	=B5-$B$15	=C5-$C$15	=D5*E5	=D5^2	=E5^2	=$D$18+$D$17*B5	=$G$18+$G$17*C5
4	2,9	6	=B6-$B$15	=C6-$C$15	=D6*E6	=D6^2	=E6^2	=$D$18+$D$17*B6	=$G$18+$G$17*C6
5	2,5	3	=B7-$B$15	=C7-$C$15	=D7*E7	=D7^2	=E7^2	=$D$18+$D$17*B7	=$G$18+$G$17*C7
6	1,4	3	=B8-$B$15	=C8-$C$15	=D8*E8	=D8^2	=E8^2	=$D$18+$D$17*B8	=$G$18+$G$17*C8
7	2,5	5	=B9-$B$15	=C9-$C$15	=D9*E9	=D9^2	=E9^2	=$D$18+$D$17*B9	=$G$18+$G$17*C9
8	2,2	3	=B10-$B$15	=C10-$C$15	=D10*E10	=D10^2	=E10^2	=$D$18+$D$17*B10	=$G$18+$G$17*C10
9	3	5	=B11-$B$15	=C11-$C$15	=D11*E11	=D11^2	=E11^2	=$D$18+$D$17*B11	=$G$18+$G$17*C11
10	3,3	5	=B12-$B$15	=C12-$C$15	=D12*E12	=D12^2	=E12^2	=$D$18+$D$17*B12	=$G$18+$G$17*C12
11	1,5	2	=B13-$B$15	=C13-$C$15	=D13*E13	=D13^2	=E13^2	=$D$18+$D$17*B13	=$G$18+$G$17*C13
Суми					=SUM(F3:F13)	=SUM(G3:G13)	=SUM(H3:H13)		
Середні	=AVERAGE(B3:B13)	=AVERAGE(C3:C13)							
	r_{xy}		=F14/SQRT(G14*H14)						
	a_1		=F14/G14		**b_1**	=F14/H14			
	a_0		=C15-D17*B15		**b_0**	=B15-G17*C15			

6. У комірках H17:H18 аналогічним способом розрахувати коефіцієнти регресії b_0 і b_1 регресійного рівняння $\tilde{X} = b_0 + b_1 \cdot Y$.

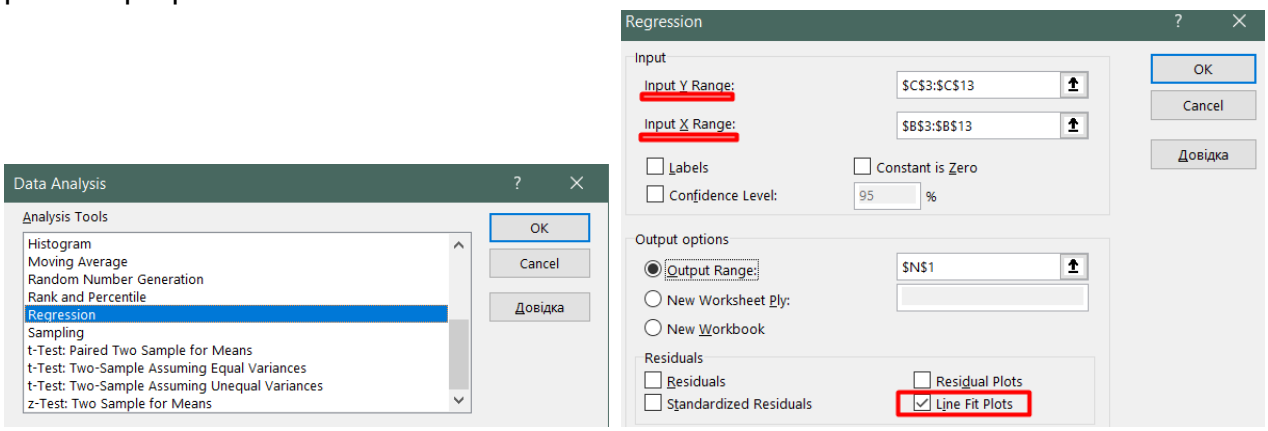
7. У комірці D21 розрахувати коефіцієнт кореляції за допомогою виразу =F14/SQRT(G14*H14) або =PEARSON(B3:B13;C3:C13), отримати $r_{xy} = 0{,}76$.

8. Побудувати графік для даних, що оброблялися та лінійну лінію тренду, встановивши позначку Показувати рівняння на діаграмі. У такий спосіб, ви можете пересвідчитися, що рохрахунки виконані коректно.

9.

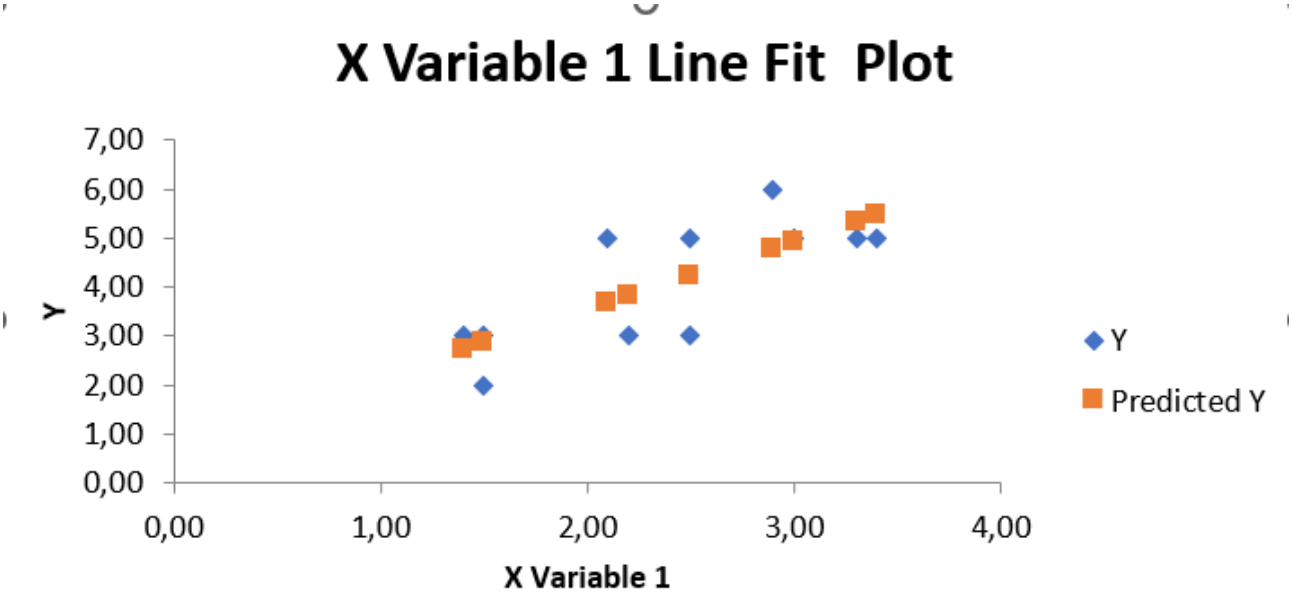


Розширення Data Analysis у MS Excel також містить функціонал для розрахунку рівнянь регресії.



Отримані рівняння регресії дають можливість аналітичного прогнозування значень залежної змінної за допомогою незалежної змінної.

	Coefficients
Intercept	0,815346882
X Variable 1	1,370007008



Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи. Розрахувати рівняння одновимірної лінійної регресії для залежностей X від Y та навпаки.

	Завдання														
1	X	2,06	2,58	3,14	3,54	4,18	4,78	5,11	5,67	6,02	6,65	7,05	7,52	8,03	8,56
	Y	14,87	15,78	16,79	18,03	18,29	19,93	20,32	21,18	2,47	23,47	24,07	25,57	27,07	27,62
2	X	2,53	3,54	3,84	3,84	4,22	4,81	6,53	5,82	6,43	7,73	8,19	7,65	9,31	9,26
	Y	19,66	20,53	21,31	22,59	23,27	24,44	25,85	26,74	27,36	28,37	29,22	30,5	31,21	32,56
3	X	2,17	2,9	3,29	4,13	5,25	4,92	5,79	5,87	6,99	7,04	8,14	8,06	8,57	9,45
	Y	12,5	13,88	15,16	16,06	16,66	17,65	18,46	19,54	20,58	21,77	22,15	23,8	24,79	25,57
4	X	3,65	3,82	3,76	5,24	5,03	5,52	5,62	6,98	6,91	7,95	7,24	9,27	8,46	10,3
	Y	10,22	10,58	12,01	12,84	13,28	15,13	15,84	17,08	17,99	18,32	19,49	20,59	21,35	23,2
5	X	3,22	3,87	4,95	5,1	5,98	7,28	6,9	7,54	7,91	8,4	8,14	8,76	9,67	10,28
	Y	16,62	17,63	19,22	19,36	20,52	21,95	22,45	23,56	24,9	25,53	26,11	28,02	28,37	29,48
6	X	2,16	2,65	3,49	3,16	3,85	4,58	5,33	5,89	6,2	6,39	6,95	7,25	7,8	8,47
	Y	15,21	15,42	16,44	17,93	18,52	19,8	20,76	21,3	22,25	24,14	24,17	25,66	26,5	27,46
7	X	4,57	5,42	5,29	6,33	7,63	7,53	7,73	8,44	9,49	9,18	10,14	9,94	10,92	11,89
	Y	12,11	12,3	13,82	14,84	15,86	16,41	17,8	18,61	19,57	21,26	21,08	22,99	23,43	24,63
8	X	2,25	2,98	2,15	2,71	3,07	4,59	4,77	5,34	5,45	6,0	6,25	6,79	8,24	8,51
	Y	16,21	17,75	16,39	18,87	19,6	21,21	21,84	23,0	24,44	25,36	25,54	27,14	27,95	28,99
9	X	6,15	5,66	7,5	6,9	8,31	8,25	9,39	9,73	9,33	10,5	11,1	11,51	12,42	12,4
	Y	10,89	11,92	12,45	13,27	14,12	15,23	16,07	17,4	18,68	19,46	20,52	21,32	22,58	23,73
10	X	1,86	1,91	2,14	3,39	3,95	4,3	5,1	5,47	5,97	6,16	6,46	6,07	6,71	7,16
	Y	7,24	8,02	9,28	10,12	11,12	12,19	13,01	14,12	15,21	16,29	17,01	18,03	19,19	20,21

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 14
МНОЖИННА ЛІНІЙНА РЕГРЕСІЯ

Мета: ознайомитися з розширеними можливостями використання табличних процесорів для визначення коефіцієнта множинної лінійної регресії.

Основні поняття: лінійна регресія, множинна регресія.

Теоретичні відомості та хід виконання роботи

Множинна лінійна регресія використовується для аналізу взаємозв'язку між залежною змінною та двома або більше незалежними змінними, при цьому враховується можливий вплив кожної змінної незалежно один від одної.

Приклад використання множинної лінійної регресії може включати аналіз впливу кількох факторів на певне явище. Наприклад, дослідження ефекту впливу витрат на рекламу (перша незалежна змінна), ціни товару (друга незалежна змінна) та економічних показників (третя незалежна змінна) на обсяг продажів товару (залежна змінна). У цьому випадку ми маємо кілька незалежних змінних, які можуть одночасно впливати на обсяг продажів товару. Множинна лінійна регресія дозволяє оцінити вагу кожної змінної та її вплив на залежну змінну при урахуванні інших змінних.

Множинна лінійна регресія – це оцінювання, наприклад, змінної Y лінійною комбінацією m незалежних змінних $X_1, X_2, \dots, X_m$.Найпростіший варіант регресії має місце для $m = 2$, коли необхідно спрогнозувати залежність однієї змінної Y від двох змінних X_1 і X_2.

Рівняння такої множинної регресії має вигляд:

$$\tilde{Y} = B_1 \cdot X_1 + B_2 \cdot X_2 + B_0$$

де $B_1 = b_1 \cdot \frac{s_y}{s_1}$, $B_2 = b_2 \cdot \frac{s_y}{s_2}$, $B_0 = \bar{Y} - B_1 \cdot \bar{X}_1 - B_2 \cdot \bar{X}_2$.

$b_1 = \frac{r_{y1} - r_{y2} \cdot r_{12}}{1 - r_{12}^2}$, $b_2 = \frac{r_{y2} - r_{y1} \cdot r_{12}}{1 - r_{12}^2}$

s_y, s_1, s_2, $\bar{Y}$, $\bar{X}_1$, $\bar{X}_2$ – стандартні відхилення і середні значення Y, X_1 і X_2;

r_{y1}, r_{y2}, r_{12} – коефіцієнти парної кореляції Пірсона між Y і X_1, Y і X_2, X_1 і X_2.

Для оцінки зв’язку, з одного боку змінної Y, а, з іншого, - двох змінних X_1 і X_2, використовують коефіцієнт множинної кореляції:

$$R = \sqrt{b_1 \cdot r_{y1} + b_2 \cdot r_{y2}}$$

У прикладі, наведеному нижче, потрібно спрогнозувати залежність змінної Y від комбінації незалежних змінних X_1 і X_2, які представлено у таблиці.

	A	B	C	D
1	**Емпіричні дані**			
2	**i**	**Y**	**X1**	**X2**
3	1	4	8	1
4	2	3	2	3
5	3	5	6	4
6	4	5	8	2
7	5	4	7	3
8	6	3	4	4
9	7	4	5	3
10	8	5	8	5
11	9	4	7	2
12	10	4	8	4
13	11	3	4	3
14	12	4	3	4

Послідовність рішення з розрахунку:

1. У комірки B15 і D15 ввести =AVERAGE(B3:B14), = AVERAGE(C3:C14) і = AVERAGE (D3:D14) й отримати середні значення масивів $\bar{Y} = 4$, $\bar{X}_1 = 5{,}83$, $\bar{X}_2 = 3{,}17$.

2. У комірки B16:D16 ввести функції =STDEV.S(B3:B14), = STDEV.S (B3:B14) і = STDEV.S (B3:B14) й отримати стандартні відхилення $s_y = 0{,}74$, $s_1 = 2{,}17$, $s_2 = 1{,}11$.

3. У комірках B17:B19 розрахувати коефіцієнти парної кореляції Пірсона за допомогою функції =PEARSON(), отримати значення $r_{y1} = 0{,}68$, $r_{y2} = 0{,}11$, $r_{12} = -0{,}21$.

4. У комірки B20 і B21 ввести вирази =(B17-B18*B19)/(1-B19^2) і =(B18-B17*B19)/(1-B19^2), отримати значення $b_1 = 0{,}74$, $b_2 = 0{,}27$.

5. У комірки E20:E22 ввести вирази =B20*B16/C16, =B21*B16/D16 і =B15-E20*C15-E21*D15. Отримати значення коефіцієнтів множинної регресії $B_1 = 0{,}25$, $B_2 = 0{,}18$, $B_0 = 1{,}97$.

6. Виконати у комірках E3:E14 розрахунки теоретичного значення $\tilde{Y}$ за рівнянням множинної регресії $\tilde{Y} = 0{,}251 \cdot X_1 + 0{,}18 \cdot X_2 + 1{,}97$. Для цього у комірку E3 ввести вираз =$E$20*C3+$E$21*D3+$E$22. Аналогічні вирази ввести в комірки E4:E14.

7. У комірку B22 ввести вираз =SQRT(B20*B17+B21*B18) і отримати значення коефіцієнта множинної кореляції $R = 0{,}73$.

Формули для розрахунків та отримані результати наведені на рисунках нижче. Отримане рівняння регресії дає можливість аналітичного прогнозування значень залежної змінної за допомогою незалежних змінних. Коефіцієнт множинної кореляції 0,73 свідчить про суттєвий прямий зв'язок між змінною Y і змінними X_1 і X_2. З іншого боку, оцінити внесок у кореляцію кожної змінної окремо не є можливим.

	A	B	C	D	E
1		**Емпіричні дані**			**Регресія**
2	**i**	**Y**	**X1**	**X2**	
3	1	4	8	1	4,16
4	2	3	2	3	3,01
5	3	5	6	4	4,19
6	4	5	8	2	4,34
7	5	4	7	3	4,26
8	6	3	4	4	3,69
9	7	4	5	3	3,76
10	8	5	8	5	4,87
11	9	4	7	2	4,09
12	10	4	8	4	4,69
13	11	3	4	3	3,51
14	12	4	3	4	3,43
15	Середні	4,00	5,83	3,17	
16	Ст.відх.	0,74	2,17	1,11	
17	r_{y1}	0,68			
18	r_{y2}	0,11			
19	$r_{1,2}$	-0,21			
20	b_1	0,74		B_1	0,25
21	b_2	0,27		B_2	0,18
22	$R_{y1,2}$	0,73		B_0	1,97

A	B	C	D	E
	Емпіричні дані			Регресія
i	Y	X1	X2	
1	4	8	1	=C3*$E$20+D3*$E$21+$E$22
2	3	2	3	=C4*$E$20+D4*$E$21+$E$22
3	5	6	4	=C5*$E$20+D5*$E$21+$E$22
4	5	8	2	=C6*$E$20+D6*$E$21+$E$22
5	4	7	3	=C7*$E$20+D7*$E$21+$E$22
6	3	4	4	=C8*$E$20+D8*$E$21+$E$22
7	4	5	3	=C9*$E$20+D9*$E$21+$E$22
8	5	8	5	=C10*$E$20+D10*$E$21+$E$22
9	4	7	2	=C11*$E$20+D11*$E$21+$E$22
10	4	8	4	=C12*$E$20+D12*$E$21+$E$22
11	3	4	3	=C13*$E$20+D13*$E$21+$E$22
12	4	3	4	=C14*$E$20+D14*$E$21+$E$22
Середні	=AVERAGE(B3:B14)	=AVERAGE(C3:C14)	=AVERAGE(D3:D14)	
Ст.відх.	=STDEV.S(B3:B14)	=STDEV.S(C3:C14)	=STDEV.S(D3:D14)	
r_{y1}	=PEARSON(B3:B14;C3:C14)			
r_{y2}	=PEARSON(B3:B14;D3:D14)			
$r_{1,2}$	=PEARSON(C3:C14;D3:D14)			
b_1	=(B17-B18*B19)/(1-B19^2)		B_1	=B20*B16/C16
b_2	=(B18-B17*B19)/(1-B19^2)		B_2	=B21*B16/D16
$R_{y1,2}$	=SQRT(B20*B17+B21*B18)		B_0	=B15-E20*C15-E21*D15

Множинна лінійна регресія допомагає розкрити складні взаємозв'язки між різними факторами та прогнозувати значення залежної змінної на основі значень декількох незалежних змінних.

Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи. Розрахувати коефіцієнт взаємної зв’язаності для заданих даних.

Варіант	Завдання

1

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	6	3,5	10	11	10	6,3	21
2	6	3,6	12	12	11	6,4	22
3	7	3,9	15	13	11	7	23
4	7	4,1	17	14	12	7,5	25
5	7	4,2	18	15	12	7,9	28
6	8	4,5	19	16	13	8,2	30
7	8	5,3	19	17	13	8,4	31
8	9	5,3	20	18	14	8,6	31
9	9	5,6	20	19	14	9,5	35
10	10	6	21	20	15	10	36

2

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	6	3,6	9	11	10	6,3	21
2	6	3,6	12	12	11	6,4	22
3	6	3,9	14	13	11	7	24
4	7	4,1	17	14	12	7,5	25
5	7	3,9	18	15	12	7,9	28
6	7	4,5	19	16	13	8,2	30
7	8	5,3	19	17	13	8	30
8	8	5,3	19	18	13	8,6	31
9	9	5,6	20	19	14	9,5	33
10	10	6,8	21	20	14	9	36

3

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	7	3,7	9	11	11	6,3	22
2	7	3,7	11	12	11	6,4	22
3	7	3,9	11	13	11	7,2	23
4	7	4,1	15	14	12	7,5	25
5	8	4,2	17	15	12	7,9	27
6	8	4,9	19	16	13	8,1	30
7	8	5,3	19	17	13	8,4	31
8	9	5,1	20	18	13	8,6	32
9	10	5,6	20	19	14	9,5	35
10	10	6,1	21	20	15	9,5	36

4

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	7	3,5	9	11	10	6,3	22
2	7	3,6	10	12	10	6,5	22
3	7	3,9	12	13	11	7,2	24
4	7	4,1	17	14	12	7,5	25
5	8	4,2	18	15	12	7,9	27
6	8	4,5	19	16	13	8,2	30
7	9	5,3	19	17	13	8,4	31
8	9	5,5	20	18	14	8,6	33
9	10	5,6	21	19	14	9,5	35
10	10	6,1	21	20	15	9,6	36

5

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	7	3,6	9	11	10	6,3	21
2	7	3,6	11	12	11	6,9	23
3	7	3,7	12	13	11	7,2	24
4	8	4,1	16	14	12	7,8	25
5	8	4,3	19	15	13	8,1	27
6	8	4,5	19	16	13	8,2	29
7	9	5,4	20	17	13	8,4	31
8	9	5,5	20	18	14	8,8	33
9	10	5,8	21	19	14	9,5	35
10	10	6,1	21	20	14	9,7	34

6

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	7	3,5	9	11	10	6,3	21
2	7	3,6	10	12	10	6,8	22
3	7	3,8	14	13	11	7,2	24
4	7	4,2	15	14	12	7,9	25
5	8	4,3	18	15	12	8,1	26
6	8	4,7	19	16	13	8,3	29
7	9	5,4	19	17	13	8,4	31
8	9	5,6	20	18	14	8,8	32
9	10	5,9	20	19	14	9,6	35
10	10	6,1	21	20	14	9,7	36

7

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	7	3,8	11	11	10	6,8	21
2	7	3,8	12	12	11	7,4	23
3	7	3,9	16	13	11	7,8	24
4	7	4,1	17	14	12	7,5	26
5	7	4,6	18	15	12	7,9	28
6	8	4,5	18	16	12	8,1	30
7	8	5,3	19	17	13	8,4	31
8	9	5,5	20	18	13	8,7	32
9	9	6,1	20	19	13	9,5	33
10	10	6,8	21	20	15	9,7	35

8

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	7	3,5	9	11	10	6,3	21
2	7	3,6	10	12	10	6,8	22
3	7	3,8	14	13	11	7,2	24
4	7	4,2	15	14	12	7,9	25
5	8	4,3	18	15	12	8,1	26
6	8	4,7	19	16	13	8,3	29
7	9	5,4	19	17	13	8,4	31
8	9	5,6	20	18	13	8,8	32
9	10	5,9	20	19	14	9,6	35
10	10	6,1	21	20	14	9,7	36

9

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	7	3,8	9	11	11	7,1	22
2	7	4,1	14	12	11	7,5	23
3	7	4,3	16	13	12	7,8	25
4	7	4,1	17	14	12	7,6	27
5	8	4,6	17	15	12	7,9	29
6	8	4,7	18	16	13	8,1	30
7	9	5,3	20	17	13	8,5	32
8	9	5,5	20	18	14	8,7	32
9	10	6,9	21	19	14	9,6	33
10	10	6,8	21	20	15	9,8	36

10

№	*y*	x_1	x_2	№	*y*	x_1	x_2
1	7	3,6	12	11	10	7,2	23
2	7	4,1	14	12	11	7,6	25
3	7	4,3	16	13	11	7,8	26
4	7	4,4	17	14	12	7,9	28
5	7	4,5	18	15	12	8,2	30
6	8	4,8	19	16	12	8,4	31
7	8	5,3	20	17	13	8,6	32
8	8	5,6	20	18	13	8,8	32
9	9	6,7	21	19	13	9,2	33
10	10	6,9	22	20	15	9,6	34

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 15
СЕРВІСИ ВЕБ-СКРЕЙПІНГУ ТА ВІДКРИТИХ ДАНИХ

Мета: ознайомитися з можливостями використання Google Spreadsheets та MS Excel для скрейпінгу даних.

Основні поняття: веб-скрейпінг, запит, таблиця, імпорт даних.

Теоретичні відомості та хід виконання роботи

Вебскрейпінг – це процес автоматичного отримання даних з веб-сторінок, що може використовуватися в наукових дослідженнях для збору і аналізу інформації з Інтернету. Вебскрейпінг може допомогти збирати великі обсяги даних з різних джерел, таких як новинні сайти, форуми, соціальні мережі тощо. Якщо веб-сторінки містять структуровану інформацію, таку як таблиці або списки, вебскрейпінг може допомогти автоматично витягнути ці дані для подальшого аналізу. а допомогою вебскрейпінгу можна аналізувати великі обсяги даних з Інтернету для виявлення глобальних тенденцій або патернів. Зібрані за допомогою вебскрейпінгу дані можна використовувати для створення датасетів, на основі яких можна проводити аналітичні дослідження.

Вебскрейпінг працює шляхом завантаження HTML-коду вебсторінки та вилучення потрібних даних із заданих елементів, таких як заголовки, ціни чи посилання. Інструменти для вебскрейпінгу можуть бути програмами (як Octoparse) або бібліотеками для мов програмування, таких як Python (BeautifulSoup, Scrapy).

Однак важливо пам’ятати про етичні та правові аспекти вебскрейпінгу, оскільки не всі сайти дозволяють автоматичне вилучення даних. Деякі сайти мають захист проти скрейпінгу (наприклад, CAPTCHA або обмеження доступу через robots.txt), тому слід дотримуватися правил і умов використання сайтів.

Проте важливо пам'ятати про етичні аспекти вебскрейпінгу, такі як дотримання правил сайту щодо обмеження доступу або використання отриманих даних. Також важливо переконатися, що використання отриманих даних відповідає принципам наукової добросовісності.

Скрейпер у Google Spreadsheets

Примітивний скрейпер міститься вже у табличному редакторі Google Spreadsheets. Тут є функція Import HTML, яка дозволяє завантажувати в документ прості таблички або списки.

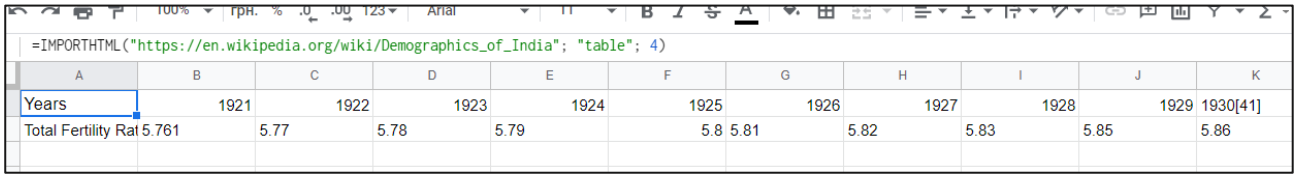
Функція скрейпінгу має вигляд:

IMPORTHTML(url, query, index),

де url — посилання на сторінку з таблицею; query — запит, може бути або table (таблиця), або list (список); index — порядковий номер таблиці чи списку на сторінці (важливо в тих випадках, коли таблиць чи списків на сторінці багато).

Таким чином, створивши таблицю в Google Spreadsheets, достатньо в комірку вставити =IMPORTHTML("https://en.wikipedia.org/wiki/Demographics_of_India"; "table"; 4) і натиснути Enter.

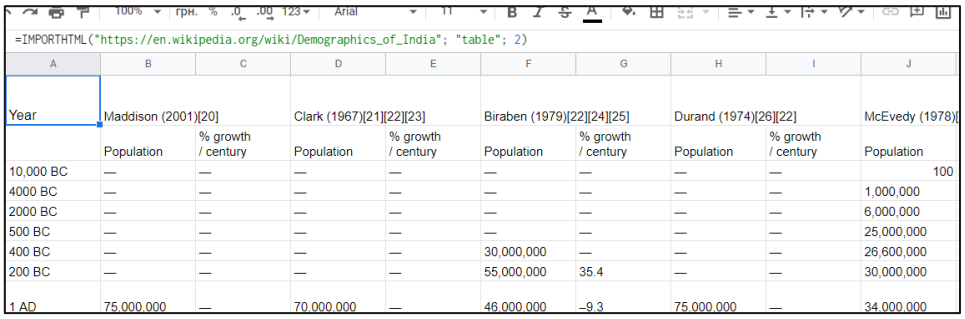
У результаті чого ми присвоюємо цій комірці значення функції IMPORTHTML із заданими параметрами, і функція витягає в нашу електронну таблицю четверту таблицю зі сторінки на вікіпедії про демографію Індії.



=IMPORTHTML("https://en.wikipedia.org/wiki/Demographics_of_India"; "table"; 4)

	A	B	C	D	E	F	G	H	I	J	K
1	Years	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930[41]
2	Total Fertility Rat	5.761	5.77	5.78	5.79	5.8	5.81	5.82	5.83	5.85	5.86

Якщо ми змінимо запит і встановивши, наприклад, номер таблиці 2, то отримаємо інший результат.



=IMPORTHTML("https://en.wikipedia.org/wiki/Demographics_of_India"; "table"; 2)

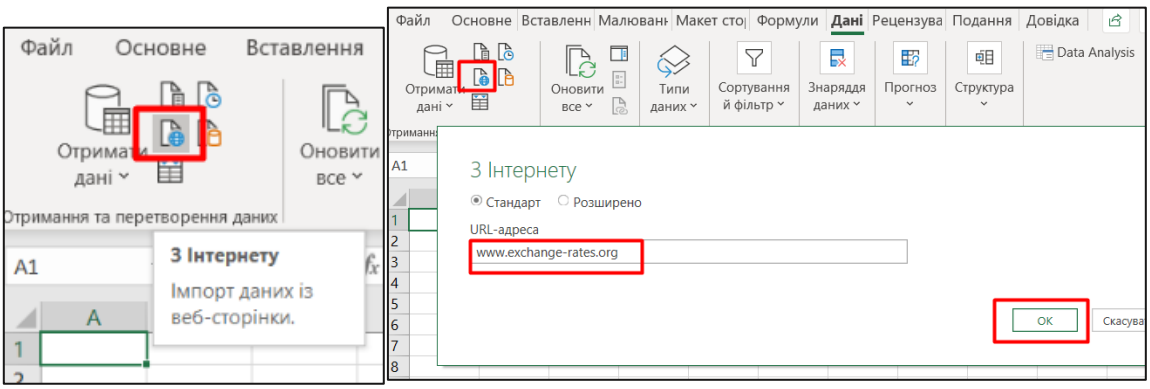
	A	B	C	D	E	F	G	H	I	J
1	Year	Maddison (2001)[20]		Clark (1967)[21][22][23]		Biraben (1979)[22][24][25]		Durand (1974)[26][22]		McEvedy (1978)[
2		Population	% growth / century	Population	% growth / century	Population	% growth / century	Population	% growth / century	Population
3	10,000 BC	—	—	—	—	—	—	—	—	100
4	4000 BC	—	—	—	—	—	—	—	—	1,000,000
5	2000 BC	—	—	—	—	—	—	—	—	6,000,000
6	500 BC	—	—	—	—	—	—	—	—	25,000,000
7	400 BC	—	—	—	—	30,000,000	—	—	—	26,600,000
8	200 BC	—	—	—	—	55,000,000	35.4	—	—	30,000,000
9	1 AD	75,000,000	—	70,000,000	—	46,000,000	−9.3	75,000,000	—	34,000,000

У випадках, коли дані на сторінці добре структуровані, функція IMPORTHTML суттєво полегшує життя і дозволяє швидко перейти власне до аналізу даних.

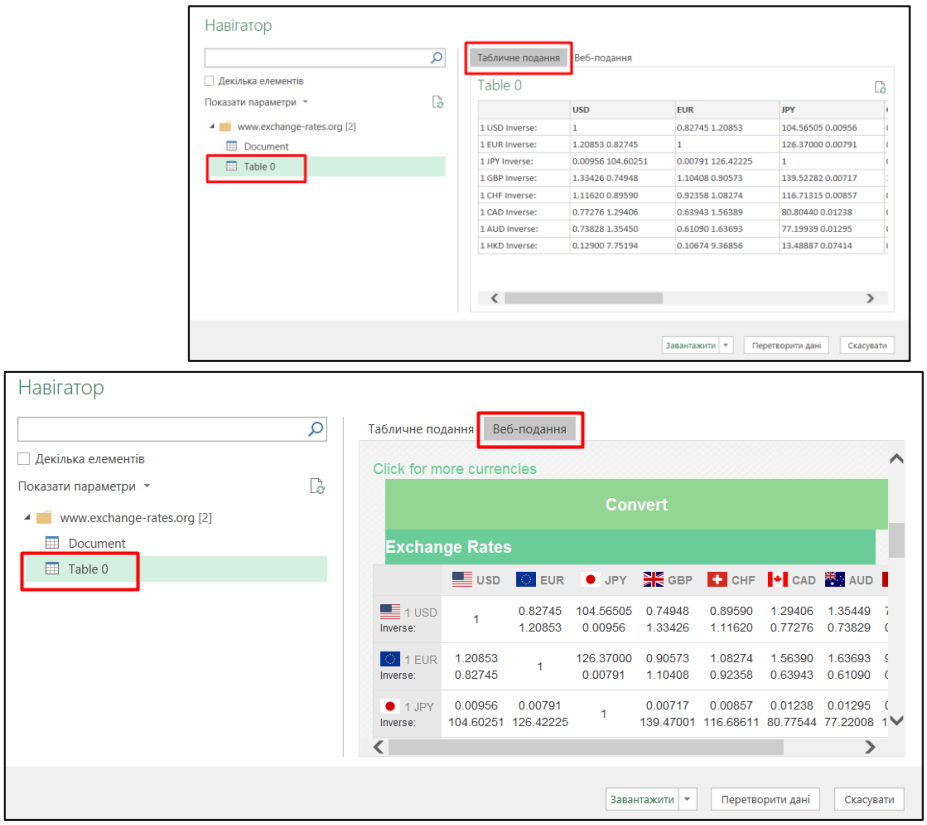
Імпорт даних за допомогою MS Excel

Імпортувати дані з веб-сторінок можна і за допомогою найпопулярнішої програми електронних таблиць — Microsoft Excel. Звісно, імпорт даних за допомогою вбудованих можливостей Excel має свої обмеження (наприклад складність роботи з багатосторінковими документами), але для деяких завдань, і для тих, хто звик працювати з програмою Excel, він може бути досить зручним.

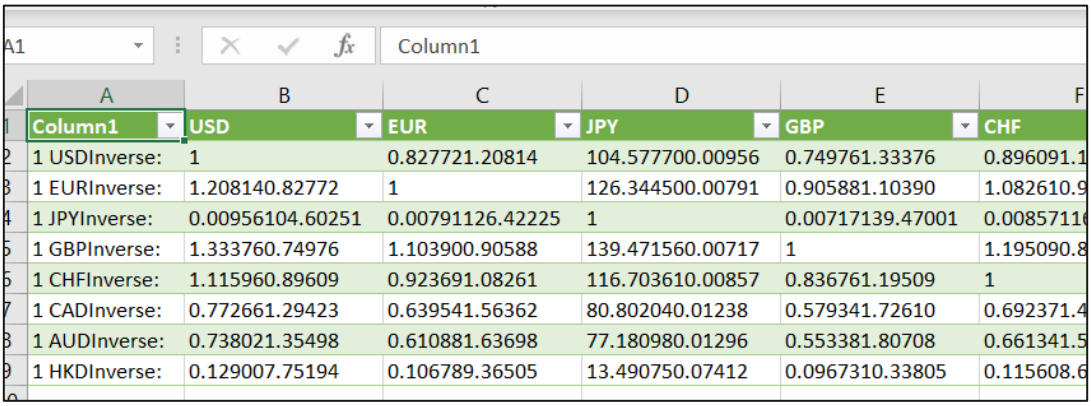
Щоб імпортувати дані в таблицю Excel, виберіть команду From Web (з вебу) в розділі Get External Data (**Отримання зовнішніх даних**) на вкладці Data (Дані). У діалоговому вікні введіть адресу веб-сайту, з якого потрібно імпортувати дані і натисніть **ОК**.



Сторінка будe завантажена у вікно для попереднього перегляду в табличному та веб-представленнях; її можна погортати і знайти потрібну інформацію.

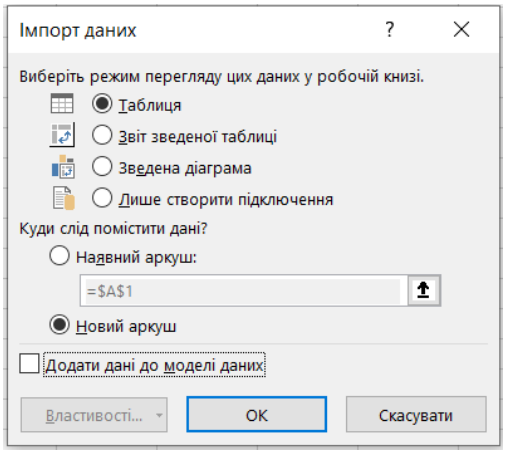


Далі, потрібно натиснути кнопку **Завантажити**.



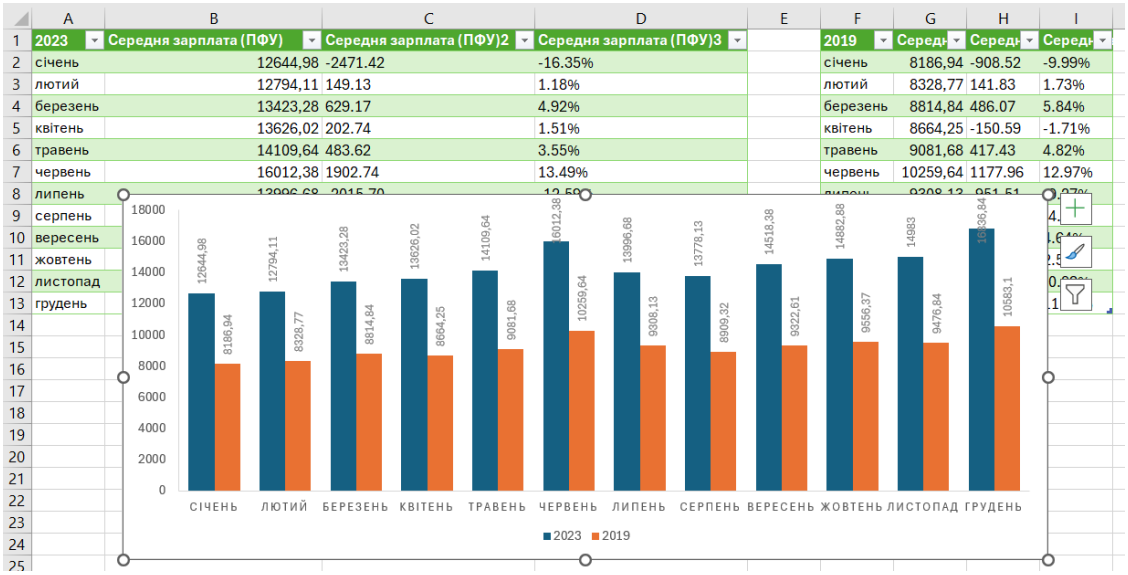
	A	B	C	D	E	F
1	Column1	USD	EUR	JPY	GBP	CHF
2	1 USDInverse:	1	0.827721.20814	104.577700.00956	0.749761.33376	0.896091.1
3	1 EURInverse:	1.208140.82772	1	126.344500.00791	0.905881.10390	1.082610.9
4	1 JPYInverse:	0.00956104.60251	0.00791126.42225	1	0.00717139.47001	0.0085711
5	1 GBPInverse:	1.333760.74976	1.103900.90588	139.471560.00717	1	1.195090.8
6	1 CHFInverse:	1.115960.89609	0.923691.08261	116.703610.00857	0.836761.19509	1
7	1 CADInverse:	0.772661.29423	0.639541.56362	80.802040.01238	0.579341.72610	0.692371.4
8	1 AUDInverse:	0.738021.35498	0.610881.63698	77.180980.01296	0.553381.80708	0.661341.5
9	1 HKDInverse:	0.129007.75194	0.106789.36505	13.490750.07412	0.0967310.33805	0.115608.6

Якщо обрати варіант **Завантажити до**, з'явиться вікно діалогу з можливістю обрати опції імпорту даних.



Імпортовані дані можна використовувати так само, як і будь-яку іншу інформацію в Excel. Їх можна використовувати для побудови графіків, спарклайнів (міні-графіків), формул. Один з плюсів імпорту даних в Excel є можливість оновлення даних прямо в самій програмі. Так, достатньо натиснути команду **Оновити** на вкладці **Дані**, і ця дія відправить запит web-сторінці і, якщо є більш свіжа версія даних, запустить процес оновлення в таблиці.

Завантажені дані можна обробляти відповідно до потреб дослідження. На рисунку наведено приклад візуалізації середньої заробітної плати в Україні (у розрізі місяців) у 2019 та 2023 роках.



Сервіси відкритих даних

У сучасному світі дані стали ключовим ресурсом для прийняття рішень у багатьох сферах, включаючи бізнес, науку, соціальні дослідження та технології. Відкриті дані — це доступні для публічного використання набори даних, які можна аналізувати, досліджувати та використовувати для розробки нових продуктів, послуг і рішень. Однією з найпопулярніших платформ для роботи з відкритими даними є Kaggle, але існують й інші важливі сервіси, що пропонують доступ до великих наборів даних.

1. Kaggle — це популярна платформа для аналітики даних і машинного навчання, де користувачі можуть змагатися у вирішенні реальних задач на основі відкритих даних. Платформа також пропонує великий репозиторій наборів даних, які можна вільно використовувати для навчання, досліджень та участі в конкурсах. Kaggle був заснований у 2010 році й сьогодні є частиною корпорації Google.

Одна з головних переваг Kaggle — це можливість швидкого доступу до якісних наборів даних, а також до інструментів для аналізу, таких як інтеграція з Python і R. Користувачі можуть використовувати Kaggle Notebooks (середовище для виконання кодів на Python і R), що дозволяє аналізувати дані прямо на платформі, не завантажуючи їх локально.

Kaggle також надає можливість брати участь у змаганнях з машинного навчання, де компанії та організації пропонують задачі з великими призовими

фондами. Це не лише допомагає фахівцям покращувати свої навички, але й дозволяє компаніям отримувати рішення для реальних проблем на основі аналізу даних.

2. Google Dataset Search — це інструмент пошуку відкритих наборів даних, який допомагає користувачам знаходити дані з різних джерел в Інтернеті. Він працює аналогічно до звичайного пошуку Google, але фокусується виключно на наборах даних. Інструмент збирає інформацію про набори даних з різних сайтів і платформ, що робить його зручним для пошуку специфічних даних для досліджень або проєктів. Джерела можуть включати урядові агентства, наукові установи або приватні організації, що публікують дані у відкритому доступі.

3. UCI Machine Learning Repository — це один з найстаріших і найбільш використовуваних репозиторіїв для машинного навчання. Він був заснований у 1987 році і містить сотні наборів даних, які використовуються для навчання алгоритмів машинного навчання та тестування їх ефективності. Набори даних включають інформацію з різних галузей, таких як біологія, медицина, фінанси та соціальні науки. UCI є чудовим ресурсом для студентів, дослідників та практиків у галузі аналізу даних і штучного інтелекту.

4. Data.gov — це портал відкритих даних уряду США, який містить понад 300,000 наборів даних. Дані охоплюють широкий спектр галузей, таких як охорона здоров'я, енергетика, транспорт, екологія та багато інших. Всі набори даних публікуються державними агентствами США і можуть бути використані для досліджень, аналізу або створення нових продуктів та послуг. Data.gov є чудовим прикладом того, як уряд може робити дані доступними для публічного використання.

5. World Bank Open Data надає доступ до величезної кількості статистичних даних, що стосуються економічного розвитку різних країн. Платформа містить набори даних про економічні показники, демографію, соціальні аспекти та екологічні фактори. Дані регулярно оновлюються та доступні для завантаження у зручних форматах для аналізу. Цей ресурс є важливим для досліджень, пов'язаних з глобальною економікою та соціальними тенденціями.

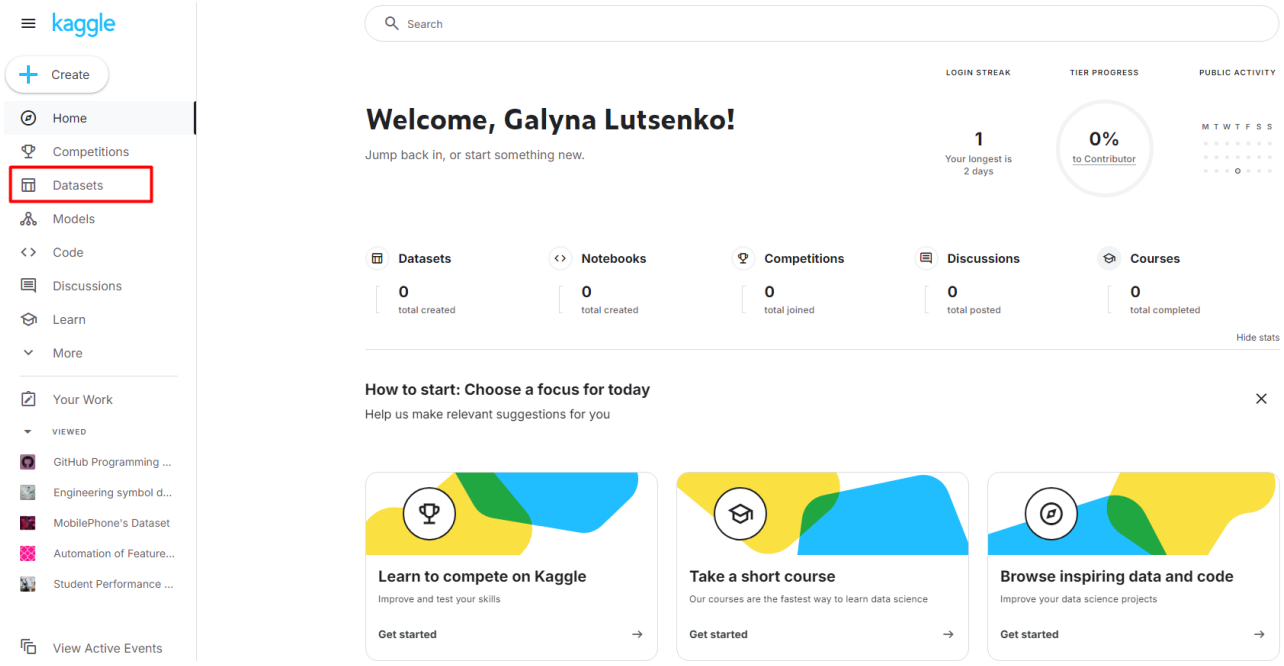
6. European Union Open Data Portal надає доступ до наборів даних, зібраних організаціями та установами Європейського Союзу. Ці дані охоплюють різні аспекти життя в ЄС, включаючи економіку, сільське господарство, науку та технології, суспільне життя та інші галузі. Портал дозволяє отримати інформацію для досліджень та аналізу, що стосується країн Європейського Союзу, і сприяє більш прозорому доступу до інформації.

7. OpenStreetMap (OSM) — це проєкт з відкритими даними, що дозволяє користувачам отримувати доступ до географічної інформації з усього світу. OSM є одним з найбільших джерел картографічних даних, які можуть бути використані для розробки геоінформаційних систем (GIS), мобільних додатків або для інших досліджень. Дані OpenStreetMap підтримуються спільнотою користувачів, що дозволяє постійно оновлювати й удосконалювати інформацію.

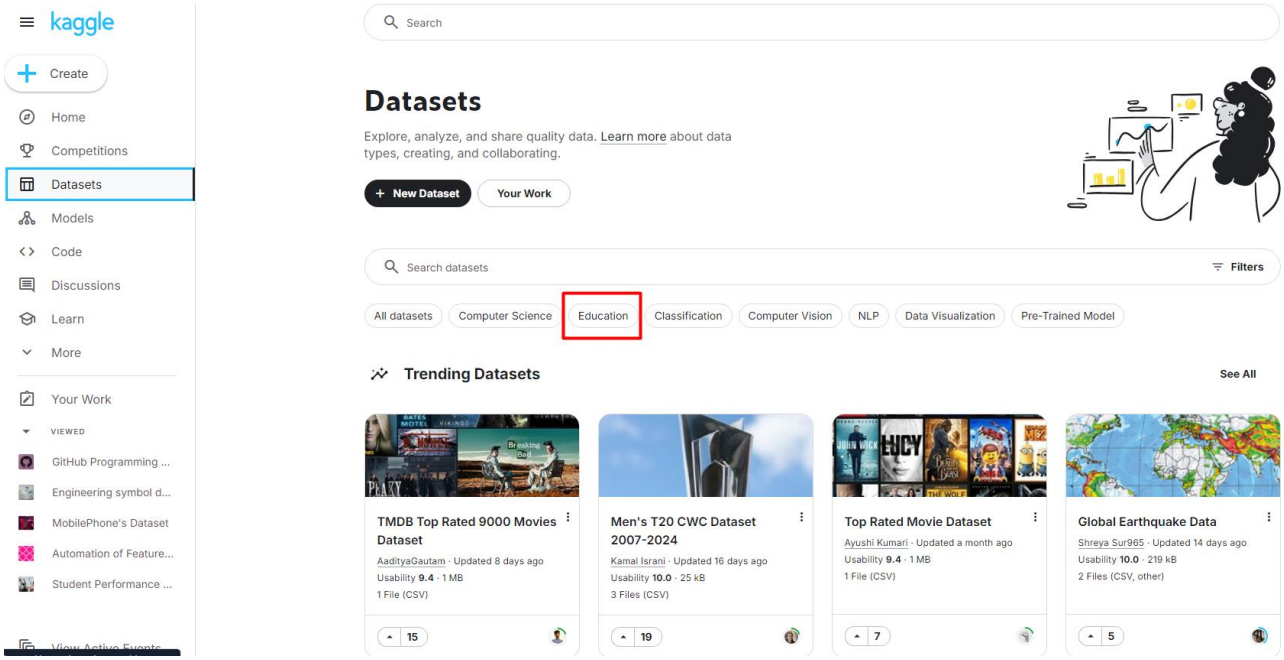
Відкриті дані мають величезне значення для розвитку науки та технологій. Вони надають дослідникам можливість отримати доступ до інформації, необхідної для проведення аналізу, тестування моделей і перевірки гіпотез. Для студентів і викладачів відкриті дані є важливим ресурсом для навчання, особливо в галузі аналізу даних, машинного навчання та статистики.

Kaggle та інші сервіси відкритих даних, такі як Google Dataset Search, UCI Machine Learning Repository, Data.gov, World Bank Open Data та інші, є важливими інструментами для аналізу даних, навчання та досліджень. Вони надають доступ до великої кількості якісних наборів даних, що дозволяє користувачам застосовувати свої знання в реальних умовах. Відкриті дані сприяють розвитку нових технологій, досліджень і підвищенню прозорості в багатьох галузях, що робить їх невід'ємною частиною сучасного світу.

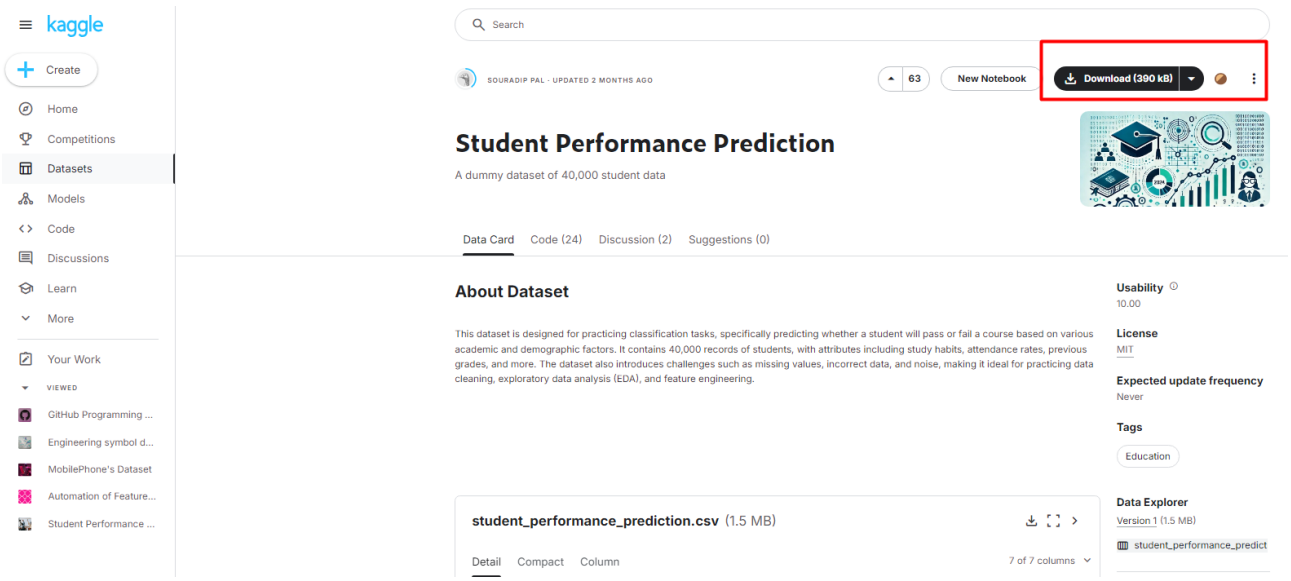
Робота з сервісом Kaggle розпочинається з реєстрації. На панелі зліва містить, зокрема, посилання на доступні в сервісі датасети.



У вікні Datasets ви можете знаходити набори даних або завантажувати власні.



Оберемо пошук наборів даних з освітньої сфери.

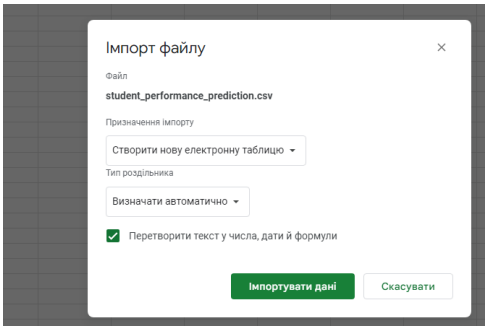


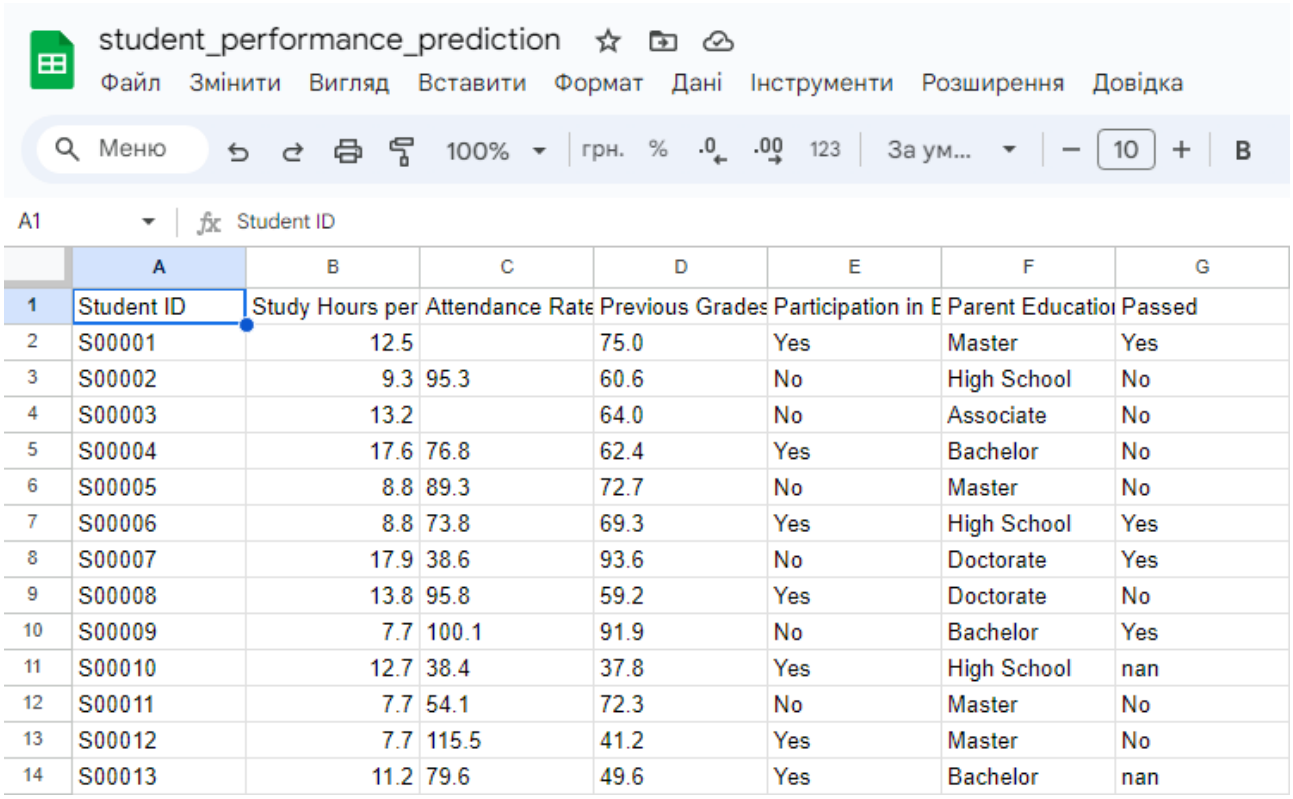
Цей набір даних призначений для відпрацювання класифікаційних завдань, зокрема прогнозування того, складе чи не здасть студент курс на основі різних академічних і демографічних факторів. Він містить 40 000 записів про студентів із такими атрибутами, як звички до навчання, рівень відвідуваності, попередні оцінки тощо.

Набір даних також містить відсутні значення, неправильні дані та шум, що робить його ідеальним для практики очищення даних, дослідницького аналізу даних (EDA) та розробки функцій.

Студентський к...	Навчальних год...	Показник відвід...	Попередні класи	Участь у позакл...	Рівень освіти ба...	Пройшов
Унікальний ідентифікатор для кожного студента (наприклад, S00001)	Середня кількість годин навчання студента на тиждень. Примітка. Цей стовпець містить деякі неправильні значення	Відсоток занять, які відвідав студент. Примітка. Цей стовпець містить деякі значення, що перевищують 100%	Середня оцінка, отримана студентом на попередніх курсах (шкала від 0 до 100). Примітка. Цей стовпець	Вказує, чи бере студент участь у позакласних заходах (Так/Ні)	Найвищий рівень освіти, отриманий батьками студента (наприклад, середня школа, молодший,	Цільова змінна, що вказує, чи склав студент курс (Так/Ні)
40000 унікальні цінності	-12.3 32.4	-14.3 150	8.3 200	немає 48% так 47% Інше (2000) 5%	Бакалавр 19% Вища школа 19% Інше (24640) 62%	так 48% немає 47% Інше (2000) 5%
S00001	12.5		75.0	так	майстер	так
S00002	9.3	95.3	60.6	немає	Вища школа	немає
S00003	13.2		64.0	немає	асоційований	немає
S00004	17.6	76.8	62.4	так	Бакалавр	немає
S00005	8.8	89.3	72.7	немає	майстер	немає
S00006	8.8	73.8	69.3	так	Вища школа	так

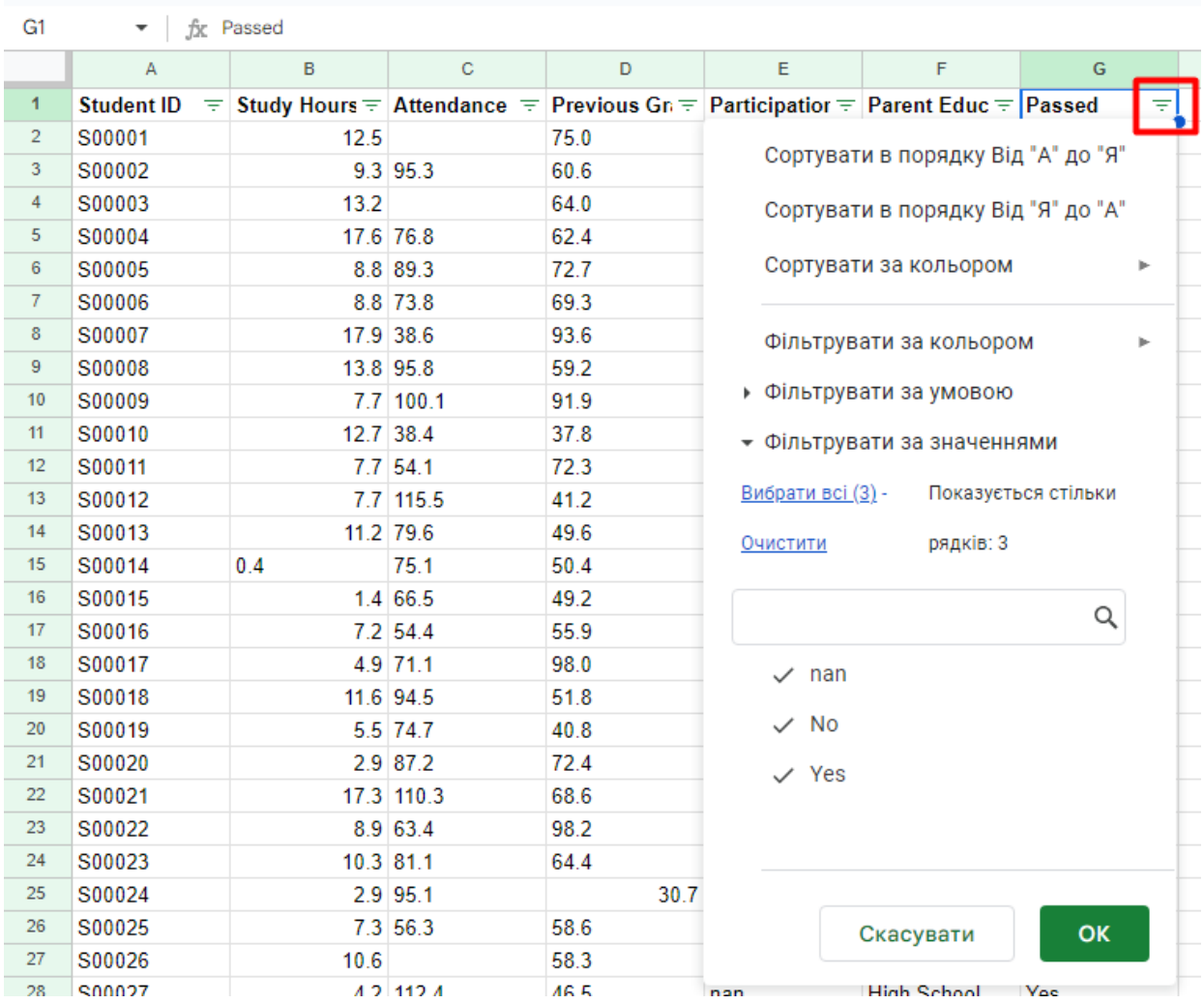
Знайдений набір даних завантажується як файл архіву. Після розархівації його можна відкрити в Google Таблицях або за допомогою MS Excel.





	A	B	C	D	E	F	G
1	Student ID	Study Hours per	Attendance Rate	Previous Grades	Participation in E	Parent Educatio	Passed
2	S00001	12.5		75.0	Yes	Master	Yes
3	S00002	9.3	95.3	60.6	No	High School	No
4	S00003	13.2		64.0	No	Associate	No
5	S00004	17.6	76.8	62.4	Yes	Bachelor	No
6	S00005	8.8	89.3	72.7	No	Master	No
7	S00006	8.8	73.8	69.3	Yes	High School	Yes
8	S00007	17.9	38.6	93.6	No	Doctorate	Yes
9	S00008	13.8	95.8	59.2	Yes	Doctorate	No
10	S00009	7.7	100.1	91.9	No	Bachelor	Yes
11	S00010	12.7	38.4	37.8	Yes	High School	nan
12	S00011	7.7	54.1	72.3	No	Master	No
13	S00012	7.7	115.5	41.2	Yes	Master	No
14	S00013	11.2	79.6	49.6	Yes	Bachelor	nan

Передані дані можна фільтрувати, групувати тощо.



	A	B	C	D	E	F	G
1	Student ID	Study Hours	Attendance	Previous Gr	Participatior	Parent Educ	Passed
2	S00001	12.5		75.0			
3	S00002	9.3	95.3	60.6			
4	S00003	13.2		64.0			
5	S00004	17.6	76.8	62.4			
6	S00005	8.8	89.3	72.7			
7	S00006	8.8	73.8	69.3			
8	S00007	17.9	38.6	93.6			
9	S00008	13.8	95.8	59.2			
10	S00009	7.7	100.1	91.9			
11	S00010	12.7	38.4	37.8			
12	S00011	7.7	54.1	72.3			
13	S00012	7.7	115.5	41.2			
14	S00013	11.2	79.6	49.6			
15	S00014	0.4	75.1	50.4			
16	S00015	1.4	66.5	49.2			
17	S00016	7.2	54.4	55.9			
18	S00017	4.9	71.1	98.0			
19	S00018	11.6	94.5	51.8			
20	S00019	5.5	74.7	40.8			
21	S00020	2.9	87.2	72.4			
22	S00021	17.3	110.3	68.6			
23	S00022	8.9	63.4	98.2			
24	S00023	10.3	81.1	64.4			
25	S00024	2.9	95.1	30.7			
26	S00025	7.3	56.3	58.6			
27	S00026	10.6		58.3			
28	S00027	4.2	112.4	46.5	nan	High School	Yes

Завдання для самостійного виконання

Завдання виконується за варіантами, що відповідають списку групи.

1. Використовуючи можливості Google Spreadsheets та/або MS Excel, знайдіть і завантажте:

1. Дані з Вікіпедії про різні країни.
2. Дані про дані про середню заробітну плату в Україні за даними Пенсійного Фонду України. Для знайдених даних побудуйте порівняльну гістограму. https://index.minfin.com.ua/ua/labour/salary/average/

Варіант	Завдання 1	Завдання 2
1	Італія	2011, 2021
2	Іспанія	2012, 2022
3	Франція	2013, 2023
4	Німеччина	2014, 2021
5	Данія	2015, 2022
6	Швейцарія	2016, 2023
7	Великобританія	2017, 2021
8	Швеція	2018, 2022
9	Норвегія	2019, 2023
10	Нідерланди	2020, 2021

2. Зареєструйтеся в сервісі Kaggle. Знайдіть та завантажте набір даних, пов’язаних з освітньою сферою. Відкрийте масив та ознайомтеся з його категоріями.

За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ЛАБОРАТОРНА РОБОТА № 16
ВІЗУАЛІЗАЦІЯ ДАНИХ. ТИПИ ГРАФІКІВ

Мета: ознайомитися з можливостями використання табличних процесорів для візуалізації даних.

Основні поняття: візуалізація даних, типи діаграм, мінідіаграми (спарклайни).

Теоретичні відомості та хід виконання роботи

Візуалізація даних є важливим аспектом аналізу, оскільки дозволяє спрощено представляти складну інформацію, що полегшує її розуміння. Google Таблиці надають потужні інструменти для створення різних типів графіків, що допомагають користувачам перетворювати числові дані на наочні зображення. Вони легко доступні, зручні у використанні та не потребують додаткових програмних знань.

Щоб створити графік у Google Таблицях, достатньо виділити дані, які потрібно візуалізувати, і вибрати відповідний тип графіку через меню "Вставка" → "Діаграма". Google Таблиці автоматично пропонують найкращий варіант візуалізації, але користувачі мають можливість змінювати тип графіку, налаштовувати стилі та параметри для кращої презентації даних.

1. Гістограма

Гістограма показує кількість даних у певних інтервалах. Це корисний інструмент для візуалізації розподілу даних і порівняння між кількома категоріями. Вона часто використовується для аналізу частоти або розподілу наборів даних.

2. Лінійний графік

Лінійні графіки відображають зміни даних за часом або іншим параметром. Вони використовуються для відстеження трендів і дозволяють бачити динаміку зміни однієї чи декількох змінних. Користувачі можуть застосовувати їх для аналізу змін у продажах, доходах або інших показниках за певний період.

3. Стовпчаста діаграма

Стовпчасті діаграми дозволяють порівнювати величини між кількома категоріями. Висота кожного стовпця відповідає значенню конкретної категорії, що робить цей тип діаграм корисним для порівняння даних, таких як фінансові показники чи результати опитувань.

4. Кругова діаграма

Кругові діаграми використовуються для представлення часток або пропорцій окремих елементів у загальному наборі даних. Вони допомагають наочно бачити, яку частку займає кожна категорія від загальної суми, наприклад, частки продажів за продуктами.

5. Графік з областями

Цей тип графіку схожий на лінійний, але підкреслює обсяг змін, заповнюючи простір під лінією. Графік з областями використовується для візуалізації кумулятивних змін у кількох категоріях або показниках з часом.

6. Радарний графік

Радарний графік застосовується для порівняння декількох показників по різним осям. Кожен показник відображається на окремій осі, а лінія, що з'єднує всі точки, дозволяє побачити загальну картину.

Коли ви обираєте тип графіку, важливо враховувати, які дані ви аналізуєте та яку інформацію хочете передати. Наприклад, для візуалізації трендів і змін у часі підходять лінійні графіки, для порівняння величин – стовпчасті, а для представлення часток – кругові.

Google Таблиці також дозволяють налаштовувати графіки, змінювати кольори, додавати мітки даних, легенди, заголовки та підписи осей, що дозволяє користувачам створювати професійні та зрозумілі візуалізації.

Розглянемо можливості, що пропонуються для візуалізації даних у табличних процесорах. Створіть на першому листі файлу таблицю, наведену нижче.

A	B	C	D	E	F	G
Витрати за перше півріччя						
	Січень	Лютий	Березень	Квітень	Травень	Червень
Продукти харчування	1730	1920	1600	1780	1720	1650
Комунальні послуги	1120	1120	1050	750	780	720
Придбання речей	1600	890	1250	2200	1230	1080
Обслуговування автомобіля	250	380	500	250	250	420
Квитки	420	1120	0	0	360	420
Інше	380	420	280	280	315	550
Щомісячні витрати	5500	5850	4680	5260	4655	4840

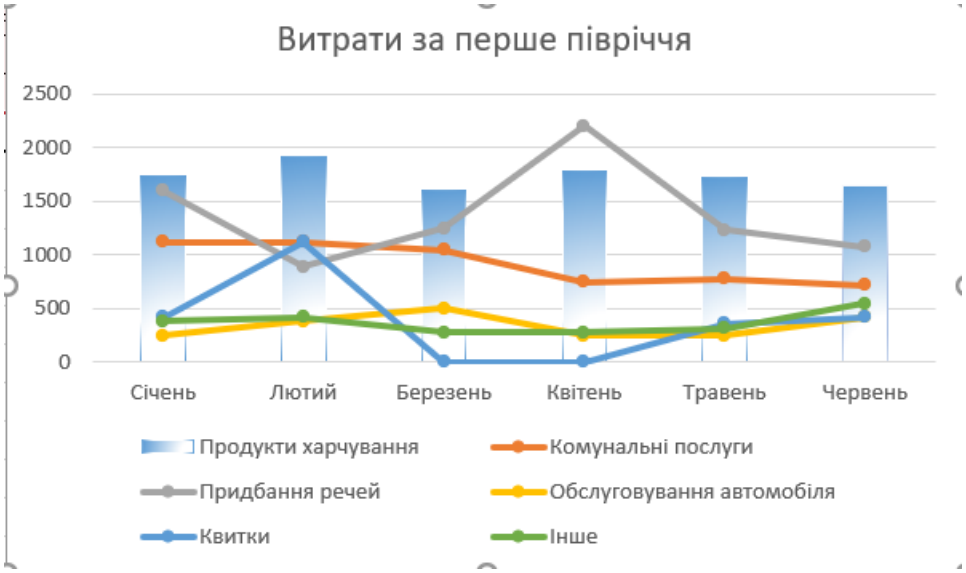
Скопіюйте таблицю на листи 2 та 3.

1. На листі 1 нижче від таблиці побудуйте базову діаграму типу Лінійчаста з маркерами. Введіть назву діаграми.

2. За необхідності (для відображення усіх надписів без спотворень) змініть розміри діаграми.

3. Змініть для ряду Продукти харчування тип діаграми на Стовпчаста діаграма з накопиченням.

4. Установіть для стовпчастої діаграми градієнтну заливку радіального типу.



5. Обрахуйте в таблиці щомісячні витрати (у відсотках, до витрат за півріччя) та на окремому листі побудуйте об’ємну секторну діаграму. Змініть кольорову гаму діаграми на власний вибір за допомогою Конструктора діаграм, який активується при обранні діаграми.

6. Встановіть підписи даних для всіх секцій.

7. На листі 2 додайте до таблиці рядок щомісячних прибутків та розрахуйте накопичення.

Використаємо спарклайни для деталізованої візуалізації трендів. Спарклайни — це мініатюрні графіки, які вбудовуються в окрему комірку та надають швидкий візуальний огляд тенденцій або змін у даних. Вони є чудовим інструментом для компактного відображення трендів, що допомагає побачити зміни в даних без необхідності створювати великі діаграми. У Google Таблицях спарклайни використовуються для відображення простих графіків прямо у комірках поруч із даними.

Типи спарклайнів у Google Таблицях:

1. Лінійний спарклайн. Лінійні спарклайни відображають зміну даних як маленьку лінію, що дозволяє бачити тренди у ряді чисел. Наприклад, такий графік може показати зростання або спад продажів за кілька місяців.

2. Стовпчастий спарклайн. Цей тип спарклайнів використовує вертикальні стовпці для представлення даних, подібно до звичайної стовпчастої діаграми, але в компактному форматі.

3. Спарклайн з областями.Спарклайн з областями заповнює простір під лінією, підкреслюючи зміну обсягу даних.

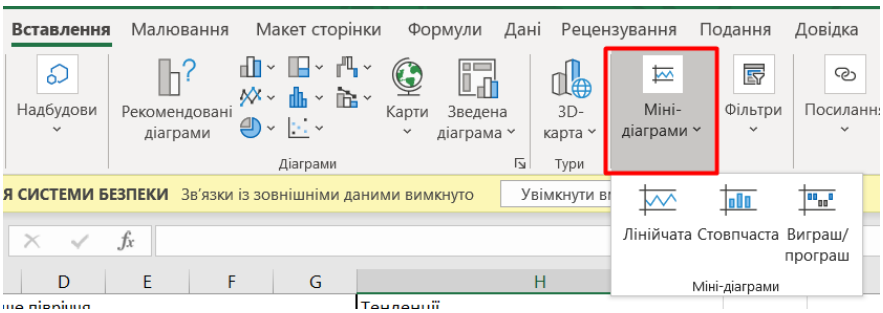
Щоб додати спарклайн, необхідно використовувати функцію SPARKLINE:

=SPARKLINE(діапазон_даних, [параметри])

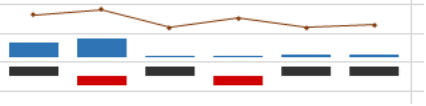
=SPARKLINE(A1:A10, {"charttype", "line"})

Ця формула створить лінійний спарклайн на основі даних із комірок A1 до A10. Спарклайни — це зручний спосіб отримати загальний огляд даних у компактній формі без потреби в великих діаграмах.

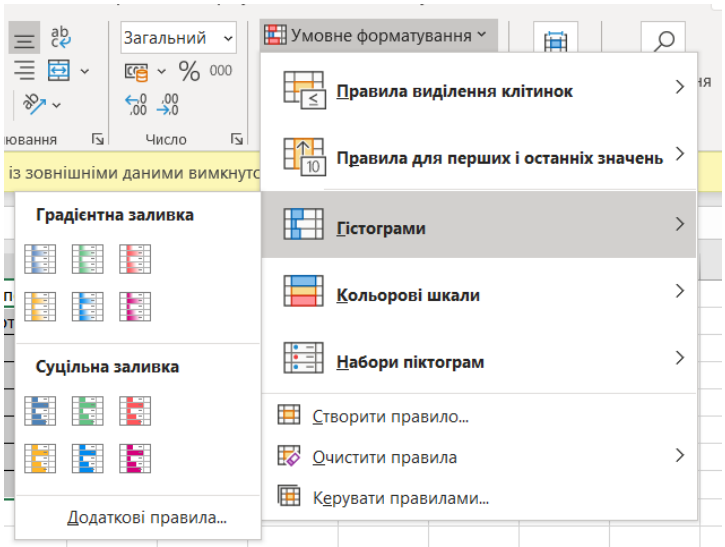
8. Додайте в таблицю стовпчик Тенденції та побудуйте в цьому стовпчику спарклайни наступних типів: для витрат – спарклайн Лінійна, для прибутків – Стовпчаста, для накопичень – спарклайн Виграш-Програш.



	Інше	380	420	280	280	315	550
	Щомісячні витрати	5500	5850	4680	5260	4655	4840
0	Прибуток	5600	5700	5200	5200	5250	5250
1	Накопичення	100	-150	520	-60	595	410



9. На листі 3 додайте стовпчик Сума та розрахуйте сумарні витрати за типами.
10. Застосуйте умовне форматування з використанням гістограм.



	Витрати за перше півріччя						
	Січень	Лютий	Березень	Квітень	Травень	Червень	Сума
Продукти харчування	1730	1920	1600	1780	1720	1650	10400
Комунальні послуги	1120	1120	1050	750	780	720	5540
Придбання речей	1600	890	1250	2200	1230	1080	8250
Обслуговування автомобіля	250	380	500	250	250	420	2050
Квитки	420	1120	0	0	360	420	2320
Інше	380	420	280	280	315	550	2225

Завдання для самостійного виконання

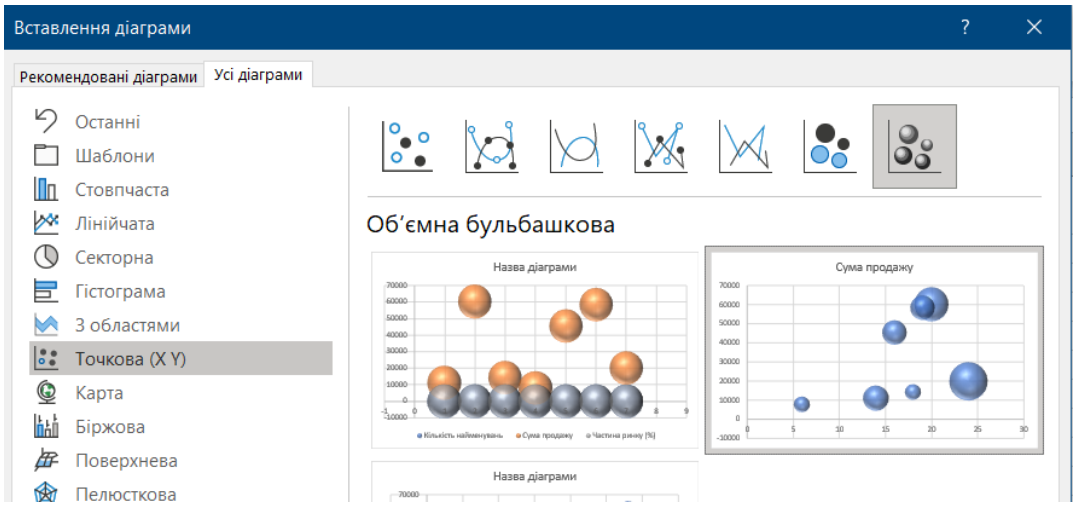
1. За наведеним зразком створіть таблицю в MS Excel чи Google Таблицях.
2. Використовуючи функцію =RANDBETWEEN(1;10) згенеруйте в стовпчиках оцінки для усіх експертів (замість значень, заданих блакитним кольором).
3. Знайдіть середню оцінку за кожною з характеристик (стовпчик G).
4. Використовуючи функцію =RANK.AVG(G3;$G$3:$G$10), визначте рейтинг середніх оцінок за характеристиками.
5. Знайдіть сумарну оцінку для кожного з експертів (рядок 11).

	A	B	C	D	E	F	G	H
1		Експерти					Середня оцінка (за характеристиками)	Ранг
2	Характеристики	1	2	3	4	5		
3	Зручність	8	5	6	4	6	5,8	3
4	Змістовність	3	6	2	8	7	5,2	3
5	Функціональність	7	1	3	1	1	2,6	4
6	Структурованість	6	3	4	2	2	3,4	7
7	Актуальність	2	3	1	3	3	2,4	6
8	Швидкодія	3	2	5	6	4	4	8
9	Дизайн	5	8	8	6	8	7	5
10	Кольорова гама	4	7	7	7	5	6	1
11	Сумарна оцінка (за експертами)	38	35	36	37	36		

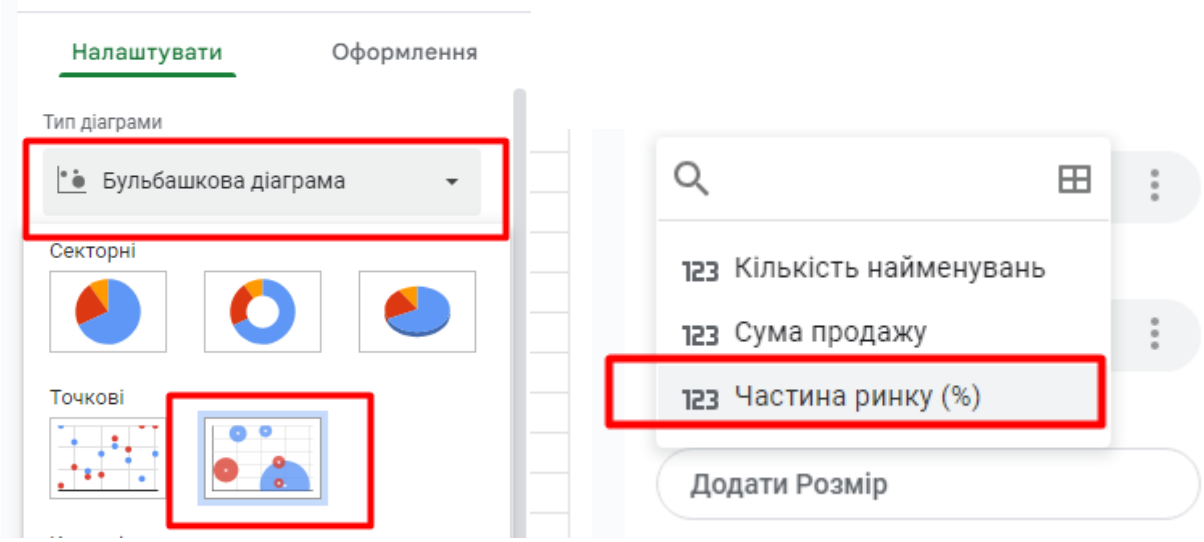
6. Побудуйте базову діаграму типу Лінійчаста з маркерами за всіма характеристиками (вісь абсцис – характеристики, вісь ординат – оцінки, ряди – номер експерта).
7. У стовпчику I побудуйте стовпчасті міні-діаграми для оцінок.
8. Побудуйте плоску стовпчасту діаграму оцінок за характеристиками від різних експертів й розташуйте її на окремому аркуші.
9. Скопіюйте таблицю і вставте її на іншому аркуші. Перегляньте можливості умовного форматування та оберіть варіант, який вам подобається.
10. Створіть таблицю, наведену нижче та побудуйте бульбашкову діаграму.

	A	B	C	D
1	Продавець	Кількість найменувань	Сума продажу	Частина ринку (%)
2	A	14	11200	13
3	B	20	60000	23
4	C	18	14400	5
5	D	6	8000	5
6	F	16	45200	12
7	G	19	58000	12
8	H	24	20000	30

MS Excel



Google Таблиці



За результатами виконання завдання сформувати звіт та завантажити в Google Classroom.

ДЖЕРЕЛА ТА РЕКОМЕНДОВАНА ЛІТЕРАТУРА

Василенко, О. А., & Сенча, І. В. (2011). *Математично-статистичні методи аналізу в прикладних дослідженнях: навчальний посібник.* Одеса: ОНАЗ ім. О.С. Попова.

Вишневецька, В. П. (2019). *Основи математичної статистики з елементами хмаро орієнтованих технологій: Лабораторний практикум для майбутніх фахівців сфери фізичної культури та спорту.* Київ: Видавництво НПУ імені М.П. Драгоманова.

Горват, А. А., Молнар, О. О., & Міньковић, В. В. (2019). *Методи обробки експериментальних даних з використанням MS Excel: Навчальний посібник.* Ужгород: Видавництво УжНУ “Говерла”.

Грицюк, М. П., & Остапчук, О. П. (2008). *Аналіз даних: Навчальний посібник.* Рівне: НУВГП.

Дьячкова О.В. (2018). *Комп'ютерний аналіз даних в MS Excel. Частина 1. Організація розрахунків і візуалізація даних.* Харків: ХНУ ім. В.Н. Каразіна.

Жалдак, М. І., Кузьміна, Н. М., & Михалін, Г. О. (2020). *Теорія ймовірностей і математична статистика: Підручник для студентів фізико-математичних та інформатичних спеціальностей педагогічних університетів. Видання четверте, доповнене.* Київ: НПУ імені М.П. Драгоманова.

Кубай Д., Газін А., Горбаль А., Шульга Є., Шаповаленко Г. Відкритий посібник з відкритих даних. Український центр суспільних даних. Київ, 2016. Режим доступу: http://socialdata.org.ua/manual/.

Peck, R., Olsen, C., & Devore, J. L. (2015). Introduction to statistics and data analysis. Cengage Learning. URL: https://www.spps.org/cms/lib/MN01910242/Centricity/Domain/859/Statistics%20Textbook.pdf

Мозгульський, Є. З., & Бородай, Г. П. (2008). *Методичні вказівки та завдання до разрахунково-графічної роботи з дисципліни "Теорія ймовірностей та математична статистика" для студентів економічних спеціальностей.* Харків: Українська державна академія залізничного транспорту.

Огірко, О. І., & Галайко, Н. В. (2017). *Теорія ймовірностей та математична статистика: навчальний посібник.* Львів: ЛьвДУВС.

Оксанич, А. П., & Рилова, Н. В. (2018). *Методичні вказівки щодо виконання лабораторних робіт з навчальної дисципліни «Обробка та аналіз даних» для аспірантів зі спеціальності 122 – «Комп’ютерні науки» (третій освітньо-професійний рівень).* Кременчук: КНУ імені М. Остроградського.

Руденко, В. М. (2012). *Математична статистика. Навчальний посібник.* Київ: Центр учбової літератури.

Сисоєва С.О., Кристопчук Т.Є. (2013). Методологія науково-педагогічних досліджень. Підручник. Рівне: Волинські обереги.

Швачич, Г. Г., Коноваленков, В. С., Соболенко, О. В., Заборова, Т. М., Христян, В. І., & Єгорцева, Є. Є. (2017). *Навчальний посібник щодо вивчення дисципліни "Методи прикладного ста-тистичного аналізу".* Діпро: НМетАУ.

Відео-курс «Data Analysis Full Course Using Statistics». URL: https://www.youtube.com/watch?v=Q-dOX4Y1fKE