

Державний вищий навчальний заклад  
«Прикарпатський національний університет імені Василя Стефаника»  
Кафедра комп'ютерних наук та інформаційних систем

**Матеріали для самостійного вивчення дисципліни**  
**«Аналіз даних»**  
**для студентів спеціальності**  
**126 «Інформаційні системи та технології»**

Матеріали для самостійного вивчення дисципліни «Аналіз даних» для студентів спеціальності 126 «Інформаційні системи та технології» – 2020. – 35 с.

Розробник:

*Ізмайлов Артем Вікторович, магістр, асистент кафедри комп'ютерних наук та інформаційних систем.*

© Ізмайлов А. В., 2020 р.

## ВСТУП

Аналіз даних – це розділ математики, що займається розробкою методів обробки даних незалежно від їх природи.

Виділяють такі етапи аналізу даних: отримання даних, обробка, аналіз та інтерпретація результатів обробки. Отримати самі дані не так важливо, як зробити на їх основі правильні висновки.

Аналіз даних можна вважати прикладним розділом математичної статистики, проте, потрібно підкреслити, що аналіз даних охоплює обробку як кількісних, так і якісних даних. При чому, не обов'язково використовуються імовірнісні моделі при описі об'єктів, явищ та процесів що досліджуються. Уміння інтерпретувати дані та робити коректні висновки на їх основі є фундаментальною компетенцією ІТ-фахівця, оскільки, дозволяє створювати ефективне програмне забезпечення шляхом обґрунтованого проектування його моделі.

Навчальна дисципліна «Аналіз даних» є однією з важливих математичних дисциплін і формує навички практичної та наукової діяльності бакалавра напряму підготовки «Інформаційні системи та технології».

**Мета** викладання навчальної дисципліни полягає в:

- Ї формуванні у студентів системного підходу до аналізу даних;
- Ї розкритті актуальних аспектів застосування сучасних інформаційних технологій з метою ефективного використання набутих теоретичних знань з даної дисципліни;
- Ї набутті студентами практичних навичок здійснення попередньої обробки даних, проведення кореляційного, дисперсійного, регресійного, коваріаційного, дискримінантного, кластерного аналізу та аналізу часових рядів, проведення факторного експерименту;
- Ї формуванні практичних навичок застосування теорії аналізу даних для розв'язування прикладних задач природничих, економічних, соціальних та інших наук;

- Ї встановленні предметних зв'язків навчальної дисципліни з різними розділами математичної статистики, фізики;
- Ї отриманні студентами теоретичної підготовки і практичних навиків для успішного засвоєння фундаментальних і спеціальних дисциплін навчального плану, а також для можливості вивчення спеціальної літератури.

**Завдання** викладання навчальної дисципліни: надати студентам інформацію щодо теоретичних та практичних аспектів аналізу даних, етапів аналізу даних, попередньої обробки даних, проведення кореляційного, дисперсійного, регресійного, коваріаційного, дискримінантного, кластерного аналізу та аналізу часових рядів, проведення факторного експерименту.

У результаті вивчення навчальної дисципліни студент повинен

**знати:**

- Ї теоретичні положення аналізу даних;
- Ї етапи аналізу даних;
- Ї визначення головних характеристик скалярних змінних;
- Ї задачу кореляційного аналізу;
- Ї задачу дисперсійного аналізу;
- Ї задачу регресійного аналізу;
- Ї задачу дискримінантного аналізу;
- Ї задачу кластерного аналізу;
- Ї задачу проведення та обробку результатів факторного експерименту;
- Ї задачу аналізу часових рядів;

**вміти:**

- Ї застосовувати математичний апарат для розв'язування задач аналізу даних;
- Ї визначати характеристики скалярних змінних;
- Ї проводити кореляційний, регресійний, дисперсійний, коваріаційний, дискримінантний, кластерний аналізи, аналізувати часові ряди та проводити обробку результатів факторного експерименту.

## **ЗМІСТ ЛЕКЦІЙНОГО КУРСУ**

### **Тема 1. Регресійний аналіз.**

- 1) Одновимірний регресійний аналіз.
- 2) Метод найменших квадратів.
- 3) Багатовимірний регресійний аналіз.
- 4) Метод послідовного виключення.

### **Тема 2. Кореляційний аналіз.**

- 1) Матриця експериментальних даних.
- 2) Загальні відомості про кореляційний аналіз.
- 3) Кореляційний аналіз кількісних ознак.

### **Тема 3. Кореляційний аналіз порядкових ознак.**

- 1) Рангова кореляція.
- 2) Коефіцієнт рангової кореляції.
- 3) Об'єднані ранги.

### **Тема 4. Факторний експеримент.**

- 1) Таблиця факторного експерименту.
- 2) Математична модель.
- 3) Визначення коефіцієнтів моделі факторного експерименту.
- 4) Частинний факторний експеримент.

### **Тема 5. Кластерний аналіз.**

- 1) Основні постановки задач класифікації.
- 2) Процес класифікації.
- 3) Кластерні алгоритми.

# ЗМІСТ ТЕМ ЛАБОРАТОРНИХ ЗАНЯТЬ КУРСУ

## Лабораторне заняття 1

### Регресійний аналіз. Метод найменших квадратів.

**Мета роботи:** Навчитись розв'язувати задачі регресійного аналізу методом найменших квадратів.

#### Хід роботи

1. Визначити (і навести обчислення) свій варіант за формулою:

$$\text{num} \bmod 4 + 1,$$

де num – номер у журналі, mod – операція взяття остачі від ділення.

2. Експериментальні дані про значення змінних  $x$  та  $y$  наведені у таблиці:

#### Варіант 1

$i$	$x_i$	$y_i$
1	0	0,2
2	2	0,4
3	3	0,5
4	4	0,7
5	5	0,9

#### Варіант 2

$i$	$x_i$	$y_i$
1	0	0,1
2	1	0,4
3	3	0,6
4	4	0,7
5	5	0,8

#### Варіант 3

$i$	$x_i$	$y_i$
1	0	2,9
2	1	3,2
3	2	3,4
4	3	3,6
5	5	3,8

#### Варіант 4

$i$	$x_i$	$y_i$
1	0	1,8
2	1	2,3
3	2	2,4
4	3	2,6
5	4	2,7

У результаті вирівнювання експериментальних даних отримана функція:

Варіант 1:

$$g(x) = \sqrt[3]{x+1} - 1,$$

Варіант 2:

$$g(x) = \sqrt[3]{x+1} - 1,$$

Варіант 3:

$$g(x) = \sqrt[3]{x+1} + 2,$$

Варіант 4:

$$g(x) = \sqrt[3]{x+1} + 1.$$

Використовуючи метод найменших квадратів апроксимувати ці дані лінійною залежністю  $y=a_0+a_1x$  (знайти параметри  $a_0, a_1$ ). Визначити, яка з двох функцій ( $g(x)$  чи  $y$ ) краще (у сенсі МНК) апроксимує експериментальні дані. Навести графік.

У процесі розв'язання задачі заповнити розрахункові таблиці:

Таблиця 1

	$x_i$	$y_i$	$x_i y_i$	$(x_i)^2$
$i = 1$				
$i = 2$				
$i = 3$				
$i = 4$				
$i = 5$				
$\sum_{i=1}^5$				

Таблиця 2

$a_0$	$a_1$

Таблиця 3

$\sigma_1$	$\sigma_2$

3. Зробити та обґрунтувати висновки стосовно того, яка функція краще апроксимує задані експериментальні дані. Зробити висновки стосовно особливостей апроксимування даних лінійними функціями.

### Контрольні запитання

1. Дайте визначення поняттю регресійного аналізу?
2. Що таке апроксимація?
3. У чому полягає метод найменших квадратів?
4. Які вимоги накладаються на експериментальні дані для задач регресійного аналізу?

## Лабораторне заняття 2

### Багатовимірний регресійний аналіз. Метод послідовного виключення.

**Мета роботи:** Навчитись розв'язувати задачі багатовимірного регресійного аналізу методом послідовного виключення.

#### Хід роботи

1. Визначити (і навести обчислення) свій варіант за формулою:

$$\text{num} \bmod 4 + 1,$$

де num – номер у журналі, mod – операція взяття остачі від ділення.

2. Експериментальні дані про значення змінних  $x_1$ ,  $x_2$  та  $y$  наведені у таблиці:

#### Варіант 1

$i$	$x_{1i}$	$x_{2i}$	$y_i$
1	1	3	9
2	4	4	10
3	9	8	12
4	11	10	14
5	12	12	15

#### Варіант 2

$i$	$x_{1i}$	$x_{2i}$	$y_i$
1	5	1	16
2	9	2	14
3	13	4	13
4	16	7	11
5	18	9	9

#### Варіант 3

$i$	$x_{1i}$	$x_{2i}$	$y_i$
1	4	8	12
2	5	9	14
3	8	13	17
4	10	15	19
5	11	19	21

#### Варіант 4

$i$	$x1_i$	$x2_i$	$y_i$
1	1	9	19
2	4	11	16
3	9	14	12
4	11	17	11
5	15	19	7

Використовуючи метод послідовного виключення апроксимувати ці дані функцією  $F(x)=c \cdot f_1(x_1) \cdot f_2(x_2)$ , де  $f_1(x_1)=a_{10}+a_{11}x_1$  та  $f_2(x_2)=a_{20}+a_{21}x_2$  (знайти параметри  $a_{10}, a_{11}, a_{20}, a_{21}$ ). Визначити, похибку апроксимації експериментальних даних. Навести тривимірний графік.

У процесі розв'язання задачі навести таблиці з усіма проміжними даними, використаними для обчислень, та заповнити розрахункові таблиці:

Таблиця 1

	$x1_i$	$x2_i$	$y_i$	$y_i^H$	$F_i$
$i = 1$					
$i = 2$					
$i = 3$					
$i = 4$					
$i = 5$					
$\sum_{i=1}^5 a$					
Середнє арифметичне					

Таблиця 2

$a_{10}$	$a_{11}$	$a_{20}$	$a_{21}$

Таблиця 3

$\sigma$

- Зробити та обґрунтувати висновки стосовно того, на скільки якісно знайдена функція  $F$  апроксимує задані експериментальні дані, вважаючи, що значення похибки апроксимації, які нижчі від одиниці, вказують на допустимий рівень

якості наближення. Зробити висновки стосовно особливостей апроксимування багатовимірних даних лінійними функціями.

### **Контрольні запитання**

1. Дайте визначення поняттю регресійного аналізу?
2. Що таке апроксимація?
3. Які особливості багатовимірної регресії слід враховувати, у порівнянні з одновимірною.
4. У чому полягає метод послідовного виключення?
5. Які вимоги накладаються на експериментальні дані для задач регресійного аналізу?

## Лабораторне заняття 3

### Кореляційний аналіз кількісних ознак

**Мета роботи:** Навчитись розв'язувати задачі кореляційного аналізу кількісних ознак на основі коефіцієнтів кореляції Пірсона та Фехнера.

#### Хід роботи

1. Визначити (і навести обчислення) свій варіант за формулою:

$$\text{num} \bmod 4 + 1,$$

де *num* – номер у журналі, *mod* – операція взяття остачі від ділення.

2. Для експериментальних даних (таблиця 1) про значення змінних *x* та *y*, які наведені у таблицях у файлі варіантів, обчислити коефіцієнт кореляції Пірсона чотирма способами:

- безпосередньо за формулою;
- за допомогою функції КОРРЕЛ MS Excel;
- за допомогою пакету MS Excel «Аналіз даних» → «Кореляція»;
- за допомогою пакету MS Excel «Аналіз даних» → «Регресія» (Множественный R) (рівень надійності виставити 0,95%).

Порівняти одержані результати. Зробити висновки про наявність та природу зв'язку між величинами *x* та *y*. Оцінити надійність отриманих результатів кореляційного та регресійного аналізу на основі показників F-критерію Фішера (F) та p-рівня (P-Значение).

3. Для експериментальних даних (таблиця 2) про значення змінних *x* та *y*, які наведені у таблицях у файлі варіантів, обчислити коефіцієнт кореляції Пірсона за допомогою пакету MS Excel «Аналіз даних» → «Регресія» (Множественный R) (рівень надійності виставити 0,95%). Побудувати графік залежності значень *y* від значень *x*. Зробити висновки про наявність та природу зв'язку між величинами *x* та *y*. Оцінити надійність отриманих результатів кореляційного та регресійного аналізу на основі показників F-критерію Фішера (F) та p-рівня (P-Значение).

4. Для експериментальних даних (таблиця 1) про значення змінних  $x$  та  $y$ , які наведені у таблицях у файлі варіантів, обчислити коефіцієнт кореляції Фехнера. Зробити висновки про наявність та природу зв'язку між величинами  $x$  та  $y$ . Порівняти результати, одержані для коефіцієнтів кореляції Пірсона та Фехнера.
5. Для експериментальних даних (таблиця 1) про значення змінних  $x$  та  $y$ , які наведені у таблицях у файлі варіантів, обчислити коваріацію двома способами:
  - безпосередньо за формулою;
  - за допомогою пакету MS Excel «Аналіз даних» → «Коваріація».
 Зробити висновки про наявність та природу зв'язку між величинами  $x$  та  $y$ . Порівняти результати, одержані для коваріації з результатами кореляційного аналізу. Вказати переваги та недоліки коваріаційного аналізу в порівнянні з кореляційним.
6. У процесі розв'язання завдань навести знімки екрану з результатами застосування пакету MS Excel «Аналіз даних» та заповнити розрахункові таблиці

Таблиця 1

№	$x_i$	$y_i$	$x_i - \bar{x}$ (1)	$y_i - \bar{y}$ (2)	(1) <sup>2</sup>	(2) <sup>2</sup>	(1)·(2)	Знак((1))	Знак((2))	Збіг знаку
1										
2										
...										
n										
$\sum_{i=1}^n$										
Сер. арифм.										

Таблиця 2

$r$	$cov$	$rF$	Кількість збігів знаку	Кількість незбігів знаку

7. Зробити висновки стосовно виявлення зв'язків між кількісними величинами на основі кореляційного аналізу.

## Експериментальні дані

### Варіант 1

Таблиця 1

$x$	$y$
3,4	14,3
3,6	14,9
4,5	17,3
4,8	17,3
4,9	17,4
5,2	17,5
5,4	17,6
5,7	17,6
6,2	17,6
6,7	17,8
7,1	18
7,5	18
7,7	18,1
7,8	18,1
7,9	18,6
8	19,7
8,4	19,8
8,5	19,9
8,8	19,9
8,9	20,1
8,9	20,1
8,9	20,1
8,9	20,1
9	20,2
9	20,3
9,1	20,3
9,3	20,5
9,4	20,6
9,7	20,9
9,7	21
9,9	21,1
10,1	21,1
10,7	21,3
11,6	21,3
12	21,8
12,3	22
12,6	22,1
12,7	22,4
13,3	22,7
13,6	23,5
13,8	24,2
14	24,4
15	25,2
15,2	25,2
15,8	25,3
15,9	25,7
16,6	26,8
17,1	27,5

Таблиця 2

$x$	$y$
-5	0
-4,795	0,566812138
-4,59	0,793160766
-4,385	0,960980749
-4,18	1,097458883
-3,975	1,213218859
-3,77	1,313748834
-3,565	1,402328064
-3,36	1,481102292
-3,155	1,551565661
-2,95	1,61480649
-2,745	1,671644699
-2,54	1,722714138
-2,335	1,768514631
-2,13	1,809446324
-1,925	1,845833145
-1,72	1,877939296
-1,515	1,905981112
-1,31	1,930135747
-1,105	1,950547615
-0,9	1,96733322
-0,695	1,980584762
-0,49	1,990372829
-0,285	1,996748357
-0,08	1,999743984
0,125	1,999374902
0,33	1,995639246
0,535	1,988518041
0,74	1,977974722
0,945	1,963954175
1,15	1,946381258
1,355	1,925158695
1,56	1,900164203
1,765	1,871246643
1,97	1,838220879
2,175	1,800860905
2,38	1,758890559
2,585	1,711970794
2,79	1,659681897
2,995	1,601498049
3,2	1,536749817
3,405	1,46456683
3,61	1,383786111
3,815	1,292796968
4,02	1,189258593
4,225	1,069532608
4,43	0,927370476
4,635	0,750122657

## Варіант 2

Таблиця 1

$x$	$y$
2,1	23,3
2,5	22,5
2,7	21,7
3	22,1
3,3	22,5
3,6	23,4
3,9	22,9
4,1	22,8
4,5	22,4
4,9	23,8
5,1	21,6
5,4	20,9
5,7	20,8
6	20,7
6,3	20,6
6,9	20,5
6,9	20,4
7,2	20,3
7,5	20,2
7,8	20,1
8,1	20
7,6	19,9
7,2	19,8
9	19,7
8,4	19,6
9,6	23,5
9,9	22,5
10,2	21,5
10,5	20,4
10,8	20,1
11,1	19,8
11,4	19,5
11,7	19,2
12	18,9
12,3	18,6
12,6	18,3
12,9	18
13,2	17,7
13,5	17,4
13,9	17,1
14,1	16,8
14,2	16,5
13,9	16,2
14,4	15,9
15,3	15,6
15,6	15,3
15,9	15
16,2	14,7

Таблиця 2

$x$	$y$
-4	0
-3,84	1,280899684
-3,68	1,353963072
-3,52	1,420399944
-3,36	1,481102292
-3,2	1,536749817
-3,04	1,58787405
-2,88	1,634899385
-2,72	1,678170432
-2,56	1,717970896
-2,4	1,754536976
-2,24	1,788067113
-2,08	1,818729227
-1,92	1,846666185
-1,76	1,872
-1,6	1,894835085
-1,44	1,915260818
-1,28	1,933353563
-1,12	1,949178288
-0,96	1,962789851
-0,8	1,974234029
-0,64	1,983548336
-0,48	1,990762668
-0,32	1,995899797
-0,16	1,998975738
0	2
0,16	1,998975738
0,32	1,995899797
0,48	1,990762668
0,64	1,983548336
0,8	1,974234029
0,96	1,962789851
1,12	1,949178288
1,28	1,933353563
1,44	1,915260818
1,6	1,894835085
1,76	1,872
1,92	1,846666185
2,08	1,818729227
2,24	1,788067113
2,4	1,754536976
2,56	1,717970896
2,72	1,678170432
2,88	1,634899385
3,04	1,58787405
3,2	1,536749817
3,36	1,481102292
3,52	1,420399944

## Варіант 3

Таблиця 1

$x$	$y$
9,7	14,3
9,6	14,9
8,9	17,3
8,5	17,3
8,1	15,6
8,2	15,1
7,3	14,1
7,9	17,6
8,2	17,6
7,8	17,8
7,4	18
7	18
6,6	18,1
6,2	18,1
5,8	18,6
5,4	19,7
5	19,8
4,6	19,9
4,2	19,9
3,8	20,1
3,4	20,1
3	20,1
2,6	20,1
2,2	20,2
1,8	20,3
1,4	20,3
1	20,5
0,6	20,6
0,2	20,9
-0,2	21
-0,6	20
-1	23
-1,4	21,3
-1,8	21,3
-2,2	21,8
-2,6	22
-3	22,1
-3,4	22,4
-3,8	28
-4,2	23,5
-4,6	24,2
-5	24,4
-4,2	25,2
-4,1	25,2
-6,2	25,3
-6,6	25,7
-7	26,8
-7,4	27,5

Таблиця 2

$x$	$y$
-6	0
-5,75	1,1426091
-5,5	1,598610508
-5,25	1,936491673
-5	2,211083194
-4,75	2,44381305
-4,5	2,645751311
-4,25	2,823512391
-4	2,98142397
-3,75	3,122498999
-3,5	3,248931448
-3,25	3,3623735
-3	3,464101615
-2,75	3,555121501
-2,5	3,636237372
-2,25	3,708099244
-2	3,771236166
-1,75	3,826080094
-1,5	3,872983346
-1,25	3,912231531
-1	3,944053189
-0,75	3,968626967
-0,5	3,986086914
-0,25	3,996526269
0	4
0,25	3,996526269
0,5	3,986086914
0,75	3,968626967
1	3,944053189
1,25	3,912231531
1,5	3,872983346
1,75	3,826080094
2	3,771236166
2,25	3,708099244
2,5	3,636237372
2,75	3,555121501
3	3,464101615
3,25	3,3623735
3,5	3,248931448
3,75	3,122498999
4	2,98142397
4,25	2,823512391
4,5	2,645751311
4,75	2,44381305
5	2,211083194
5,25	1,936491673
5,5	1,598610508
5,75	1,1426091

## Варіант 4

Таблиця 1

$x$	$y$
4,6	33,8
5,8	33,2
8,9	32,6
4	32
3,8	31,4
3,6	30,8
3,4	30,2
3,2	29,6
3	29
2,8	28,4
2,6	27,8
2,4	27,2
2,2	26,6
2	26
1,8	25,4
1,6	24,8
1,4	24,2
1,2	23,6
1	23
0,8	22,4
0,6	21,8
0,4	21,2
0,2	20,6
0,1	20
-0,2	19,4
-0,4	18,8
-0,6	22,5
-0,8	21,5
-1	28,5
-1,2	29,6
-1,4	13,4
-1,6	19,5
-1,8	18,7
-2	17,9
-2,2	17,1
-2,4	16,3
-2,6	15,5
-2,8	14,7
-3	13,9
-3,2	13,1
-3,4	12,3
-3,6	11,5
-3,8	10,7
-4	9,9
-4,2	9,1
-4,4	8,3
-4,6	7,5
-4,8	6,7

Таблиця 2

$x$	$y$
-7	0
-6,71	0,854556813
-6,42	1,195675883
-6,13	1,448489009
-5,84	1,653997014
-5,55	1,828222623
-5,26	1,979443335
-4,97	2,112605027
-4,68	2,230943792
-4,39	2,336712762
-4,1	2,431552792
-3,81	2,516699734
-3,52	2,593108606
-3,23	2,661532089
-2,94	2,722572313
-2,65	2,776716237
-2,36	2,824360502
-2,07	2,865829296
-1,78	2,901387423
-1,49	2,931249995
-1,2	2,955589654
-0,91	2,974541982
-0,62	2,988209484
-0,33	2,996664472
-0,04	2,99995102
0,25	2,998086124
0,54	2,991060149
0,83	2,978836576
1,12	2,961351043
1,41	2,938509618
1,7	2,910186192
1,99	2,876218819
2,28	2,836404738
2,57	2,790493702
2,86	2,738179046
3,15	2,679085665
3,44	2,612753649
3,73	2,538615644
4,02	2,455964834
4,31	2,363908472
4,6	2,261298164
4,89	2,14662098
5,18	2,017820606
5,47	1,871984105
5,76	1,704744879
6,05	1,509003591
6,34	1,271666109
6,63	0,962436322

### **Контрольні запитання**

1. Дайте визначення поняттю кореляційного аналізу?
2. Що таке апроксимація?
3. Як визначають коефіцієнт кореляції Пірсона?
4. Як визначають коефіцієнт кореляції Фехнера?
5. Як визначають коваріацію двох величин?

## Лабораторне заняття 4

### Кореляційний аналіз порядкових ознак

**Мета роботи:** Навчитись розв'язувати задачі кореляційного аналізу порядкових ознак на основі коефіцієнта рангової кореляції.

#### Хід роботи

1. Визначити (і навести обчислення) свій варіант за формулою:

$$\text{num} \bmod 4 + 1,$$

де num – номер у журналі, mod – операція взяття остачі від ділення.

2. Для експериментальних даних (таблиця 1) про оцінки учнів з фізики та музики, провести ранжування успішності за обома предметами та здійснити дослідження зв'язку успішності з одного предмету від успішності з іншого на основі коефіцієнта рангової кореляції. Обчислення коефіцієнту рангової кореляції здійснити за формулою (1) попередньо приписавши значення кожній парі об'єктів.

У процесі розв'язання завдання заповнити розрахункові таблиці:

Таблиця 1

Присвоєні ранги		
Учні	Фізика	Музика
А		
...		
І		

Таблиця 2

Пара	Приписане значення
АВ	
...	
ІІ	

Таблиця 3

$P$	$Q$	$S$	$n$	Кількість пар	$\tau$

Проаналізувати одержані результати. Зробити висновки про наявність та природу зв'язку між заданими величинами.

- Для утвореного у попередньому завданні ранжування здійснити дослідження зв'язку успішності з одного предмету від успішності з іншого на основі коефіцієнта рангової кореляції. Обчислення коефіцієнту рангової кореляції здійснити за однією із формул (2), крім першої, попередньо упорядкувавши ранги об'єктів за одним із предметів (наприклад, фізикою). Навести обчислення числа  $P$  або  $Q$  (залежно від обраної формули) у вигляді суми (виписати усі доданки, з яких вони утворені).

У процесі розв'язання завдання заповнити розрахункові таблиці:

Таблиця 4

Упорядковані ранги		
Учні	Фізика	Музика
	1	
	...	
	10	

Таблиця 5

$P$ (або $Q$ )	$n$	Кількість пар	$\tau$

Проаналізувати одержані результати. Порівняти обчислене значення коефіцієнту кореляції зі значенням, обчисленим у попередньому завданні. Зробити висновки про зручність обчислення коефіцієнту рангової кореляції різними способами.

- Для експериментальних даних (таблиця 2) про оцінки учнів з фізики та музики, провести ранжування успішності за обома предметами та здійснити дослідження зв'язку успішності з одного предмету від успішності з іншого на основі коефіцієнта рангової кореляції. Обчислення коефіцієнту рангової

кореляції здійснити за формулою (3). При цьому навести усі додаткові обчислення і таблиці, наприклад, упорядкованих рангів.

У процесі розв'язання завдання заповнити розрахункові таблиці:

Таблиця 6

Присвоєні ранги		
Учні	Фізика	Музика
A		
...		
R		

Таблиця 7

$P$	$Q$	$S$	$n$	$T$	$U$	Кількість пар	$\tau$

Проаналізувати одержані результати. Зробити висновки про наявність та природу зв'язку між заданими величинами. Зробити висновки про вплив об'єднаних рангів на обчислення коефіцієнту рангової кореляції та дослідження зв'язків між порядковими величинами.

- Зробити висновки стосовно виявлення зв'язків між порядковими величинами на основі кореляційного аналізу.

### Експериментальні дані

#### Варіант 1

Таблиця 1

Учні	Фізика	Музика
A	7	5
B	11	4
C	5	11
D	2	12
E	12	10
F	4	7
G	3	9
H	9	8
I	6	6
J	8	2

Таблиця 2

<i>Учні</i>	<i>Фізика</i>	<i>Музика</i>
A	12	6
B	5	12
C	11	12
D	6	8
E	11	9
F	11	11
G	10	9
H	7	12
I	11	12
J	8	7
K	7	8
L	9	5
M	10	6
N	11	8
O	6	9
P	4	10
Q	5	6
R	11	12

## Варіант 2

Таблиця 1

<i>Учні</i>	<i>Фізика</i>	<i>Музика</i>
A	5	9
B	12	6
C	11	12
D	6	4
E	8	10
F	7	5
G	9	11
H	4	8
I	2	7
J	1	2

Таблиця 2

<i>Учні</i>	<i>Фізика</i>	<i>Музика</i>
A	5	12
B	12	9
C	11	7
D	9	6
E	8	10
F	9	10
G	12	12
H	6	4
I	7	6
J	7	10
K	11	11
L	11	6
M	12	5
N	10	8
O	8	9
P	4	6
Q	8	10
R	12	6

### Варіант 3

Таблиця 1

<i>Учні</i>	<i>Фізика</i>	<i>Музика</i>
A	9	6
B	6	12
C	10	10
D	11	7
E	7	11
F	12	9
G	6	8
H	1	3
I	4	4
J	3	2

Таблиця 2

<i>Учні</i>	<i>Фізика</i>	<i>Музика</i>
A	7	12
B	10	9
C	9	9
D	7	8
E	6	11
F	6	12
G	12	7
H	9	11
I	12	6
J	5	5
K	7	9
L	8	5
M	12	4
N	11	10
O	10	11
P	7	5
Q	11	12
R	4	10

## Варіант 4

Таблиця 1

<i>Учні</i>	<i>Фізика</i>	<i>Музика</i>
A	6	2
B	10	9
C	4	12
D	12	7
E	11	10
F	7	11
G	9	8
H	5	5
I	8	4
J	2	6

Таблиця 2

<i>Учні</i>	<i>Фізика</i>	<i>Музика</i>
A	9	12
B	9	5
C	6	11
D	12	10
E	11	11
F	7	9
G	12	7
H	10	8
I	9	7
J	4	9
K	5	5
L	11	7
M	9	11
N	5	7
O	6	9
P	3	6
Q	12	9
R	3	11

### Контрольні запитання

1. Дайте визначення поняттю кореляційного аналізу?
2. Як відбувається процедура ранжування за кількісною ознакою?
3. Як визначають коефіцієнт рангової кореляції?
4. Що таке об'єднані ранги і які причини їх застосування?
5. Як визначають коефіцієнт рангової кореляції у випадку наявності об'єднаних рангів?

## Лабораторне заняття 5

### Кластерний аналіз. Ієрархічні агрегаційні алгоритми.

**Мета роботи:** Навчитись розв'язувати задачі кластерного аналізу за допомогою ієрархічних агрегаційних алгоритмів.

#### Хід роботи

Постановка задачі:

Для семи марок автомобілів здійснено оцінку за 10-бальною шкалою за рядом характеристик (таблиця 1). Необхідно здійснити кластеризацію оцінених марок за допомогою різних алгоритмів з метою подальшого створення системи автоматичного надання рекомендацій покупцям щодо купівлі автомобіля. Результати кластеризації подати у вигляді дендрограм та висновків на їх основі.

Таблиця 1 – Оцінка характеристик марок автомобілів

Марки	Маса	Ціна	Макс. швидкість	Споживання палива	Безпека
Марка 1	5	2	7	4	4
Марка 2	4	1	4	3	3
Марка 3	5	1	6	4	4
Марка 4	6	7	9	6	8
Марка 5	6	7	10	6	8
Марка 6	9	8	5	7	2
Марка 7	10	10	3	10	3

1. Обчисліть та включіть у звіт матриці відстаней між об'єктами (марками) на основі  $L_1$  та  $L_2$ -норм. Зробіть висновки щодо відмінностей між даними у отриманих матрицях та щодо того, на основі якої із них доцільно проводити побудову кластерів.
2. На основі евклідової відстані у просторі  $L_2$  ( $L_2$ -норми) здійсніть кластеризацію марок автомобілів за даними із таблиці 1 на основі усіх алгоритмів кластеризації із таблиці 2. Результати кластеризації оформити у вигляді відповідних дендрограм, які необхідно включити у звіт.

Для алгоритмів одинарного зв'язку (Single linkage), повного зв'язку (Complete linkage) та середнього зваженого зв'язку (Average linkage (weighted)) включити у звіт усі проміжні обчислення (у вигляді відповідних матриць відстаней), які здійснені у процесі побудови кластерів.

Таблиця 2 –Алгоритми кластеризації

Назва
Одинарний зв'язок (Single linkage)
Повний зв'язок (Complete linkage)
Середній зв'язок (незважений) (Average linkage (unweighted))
Середній зв'язок (зважений) (Average linkage (weighted))
Центроїд (Centroid)
Уорд (Ward)

- Зробити висновки стосовно результатів кластеризації на основі кожного із алгоритмів. Здійснити порівняння результатів та на його основі рекомендувати один із застосованих алгоритмів у якості основного для розв'язання поставленого завдання.

### Контрольні запитання

- Дайте визначення поняттю кластерного аналізу?
- Як відбувається процедура побудови матриці відстаней для неперервних величин?
- Як відбувається процедура побудови матриці відстаней для бінарних величин?
- Які методи кластеризації Ви знаєте?
- Які Ви знаєте алгоритми кластеризації? Наведіть приклади.

## **Рекомендована література**

### **Базова література**

1. Барсегян А. А., Куприянов М. С., Ступаненко В. В., Холод И. И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.: ил.
2. Гайдышев И. Анализ и обработка данных: специальный справочник – СПб.: Питер, 2001. – 752 с.: ил.
3. Дж. Тьюки. Анализ результатов наблюдений. Разведочный анализ. – М.: Мир. – 1981. – 693 с.
4. Факторный, дискриминантный и кластерный анализ: Пер. с англ./Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др.; Под ред. И. С. Енюкова. – М.: Финансы и статистика, 1989. – 215 с.: ил.
5. Мандель И. Д. Кластерный анализ. – М.: Финансы и статистика. 1988. – 176 с.: ил.
6. Иберла К. Факторный анализ / Пер с нем. В. М. Ивановой; Предисл. А. М. Дурброва. – М.: Статистика, 1980. – 398 с., ил.

### **Допоміжна література**

1. Кендел М. Ранговые корреляции. – Зарубежные статистические исследования. М., «Статистика», 1975
2. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: в 2-х кн. Кн. 1/Пер. с англ. – 2-е изд., перераб и доп. – М.: Финансы и статистика, 1986.–366 с., ил.
3. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: в 2-х кн. Кн. 2/Пер. с англ. – 2-е изд., перераб и доп. – М.: Финансы и статистика, 1987.–351 с., ил.
4. Справочник по прикладной статистике. В 2-х т. Т.1: Пер. с англ. / Под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. – М.: Финансы и статистика, 1989. – 510 с., ил.
5. Справочник по прикладной статистике. В 2-х т. Т.1: Пер. с англ. / Под ред. Э. Ллойда, У. Ледермана, С. А. Айвазяна, Ю. Н. Тюрина. – М.: Финансы и статистика, 1990. – 526 с., ил.

### **Інформаційні ресурси**

[www.scientific-library.net](http://www.scientific-library.net) – Електронна бібліотека науково-технічної літератури

[www.elibrary.ru](http://www.elibrary.ru) – Наукова електронна бібліотека науково-технічної літератури

[www.coursera.org](http://www.coursera.org) – Електронна база курсів