До уваги студентів спеціальності «Транспортні технології (автомобільний транспорт)», І курс (скорочений термін навчання), група 1906.

1. Прошу надсилати виконані завдання за адресою: wnyrk15@gmail.com З цих завдань складатиметься оцінка за навчальну роботу.

2. Студентам, які не пройшли тестування за модулем 1, за модулем 2 та з підсумкового контролю (залік)– буде виставлено оцінку «незараховано».

3. Дату заліку назначаю на 12 травня 2020 р. (термін 1год. 20хв., тобто одна пара), початок – 2 пара в 10 год. 05хв. (до 11 год.25 хв.)

4. Залік будемо проводити на ЕНК ВМ Т. Підсумковий тест з дисципліни "Вища математика", 4 семестр (30 тестів по 1 балу на протязі 1год. 20 хв.) Відомість маю закрити наступного дня (див. положення на сайті НУБіП).

Доц. Панталієнко Л.А.

Завдання №8. Знайти коефіцієнт кореляції двох випадкових величин X та Y.

№ спостереження	Вар.1		Вар.2		Вар.3		Вар.4		Вар.5		Вар.6	
	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
1	15	20	30	20	1	10	21	1	1,1	10	20	1
2	16,1	22	19	31	3	11	20	0,1	1,2	8	21	1,1
3	18	21	21	32	2	9	21	0,8	1,8	9	22	1
4	17	28	18	31	4	8	22	-0,1	2	11	19	1,3
5	11	29	17	30	5	7	19	0,7	-2	12	18	1,8
6	12,2	27	16	33	6	10	25	0,6	-1	9	21	2
7	13	26	15	35	7	12	20	0,5	2	8	19	1,3
8	14,4	28	18	38	8	15	14	0,4	1,3	22	1	1,5
9	15,8	30	19	40	9	10	17	0,9	1,6	14	25	1,2
10	16	31	20	35	8	18	25	1,1	1,8	16	17	1,4

№ спостереження	Вар.7		Вар.8		Вар.9		Вар.10		Вар.11		Вар.12	
	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
1	25	42	-5	2,6	20	18	1,5	6	55	56	0,1	0
2	30	38	7	2,3	20	19	1,1	3	58	56	0,2	0,1

3	35	24	10	1	30	23	0,7	2	60	56	0,4	0,4
4	39	22	12	1,8	19	18	0,2	0,9	59	67	0,3	0,3
5	40	21	15	1,4	29	24	-0,1	2	58	53	0,4	0,6
6	44	21	20	1	23	20	-0,2	4	57	51	0,5	0,2
7	24	31	20	1,1	24	23	-1	0,8	50	50	0,6	0,1
8	39	21	28	1,1	24	22	-1,2	1,3	56	50	0,8	0,3
9	31	30	30	1,2	30	20	-2	2	60	60	1	0,4
10	31	32	31	3	29	20	-1	1,9	58	58	0,8	0,6

№ спостереження	Вар.13		Вар.14		Вар.15		Вар.16		Вар.17		Вар.18	
	X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
1	1,2	5	0,9	9	5	0,1	10	25	13	0,5	11	19
2	1	4	2	10	4	0,5	20	30	12	-0,2	13	20
3	0,8	2	1,8	9	8	0,3	29	26	10	0,3	15	19
4	0,3	1	1,3	7	7	0,4	51	29	17	0,7	18	18
5	-0,1	2	1,4	6	10	0,8	10	31	18	0,9	13	17
6	-0,3	3	1,3	9	12	1	19	35	19	1	19	20
7	-1	0,7	1,6	11	9	1,2	29	31	14	1,3	20	21
8	-1,3	1	1,7	14	8	1,1	49	41	11	2,1	22	23
9	-2	2	1,8	9	10	1,6	51	40	10	2	24	28
10	-1	1,5	1,7	12	14	2	50	35	9	-0,8	25	30

№ спостереження	Вар.19		Вар.20		Вар.21		Вар.22		Вар.23		Вар.24	
	X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
1	-1	1,1	1,1	2,9	0	1,3	10	8	1,2	0	2	1
2	1,1	1,3	1,2	3,1	2,7	-1,9	11	10	1,3	-0,2	5	7
3	1,1	1,5	1,2	3,1	0,2	1	13	6	1,6	0,8	3	0
4	1,2	1,6	1,3	3,2	2,4	1,5	12	6	1,4	0,6	7	2
5	1,8	1	1,4	3	0,7	1	14	5	1,8	-0,1	10	5
6	1,7	1,5	1,7	3	2	-1	10	7	1,7	0,7	11	4
7	1,6	1,6	1,6	3,1	1,1	0,6	15	8	1,1	0,5	12	1
8	1,7	1,5	1,8	3	1,4	0,1	20	5	1	0,4	9	8
9	1,6	1,6	2	2,8	1,7	-0,5	19	8	1,2	0,3	8	6
10	1,7	1,2	1,9	2	1,8	-0,2	18	2	1,6	0	7	4

№ спостереження	Вар.25		Вар.26		Вар.27		Вар.28		Вар.29		Вар.30	
	X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
1	1	12	21	20	8	1	22	5	10	0,1	8	3
2	3	13	20	19	7	2	28	8	11	0,7	7	2
3	5	18	19	17	9	5	23	11	13	0,8	6	1
4	8	10	21	18	8	7	27	4	9	0,9	5	2
5	7	11	23	16	6	8	20	7	8	1	3	4

6	6	17	25	15	5	6	19	8	7	1,1	2	7
7	5	12	28	14	4	2	18	3	10	1,2	2	8
8	3	13	19	20	3	5	19	4	11	1,4	4	9
9	2	16	20	21	2	7	20	5	14	1,3	5	10
10	8	15	21	22	1	6	21	2	15	1,7	8	8

Теоретичний матеріал до завдання №8 і до заліку.

7.1. Залежність між випадковими величинами. Вибіркові рівняння регресії.

У багатьох наукових дослідженнях виникає необхідність проводити одночасно спостереження над кількома випадковими величинами, щоб встановити та оцінити їх взаємозв'язок.

Дві випадкові величини X та Y можуть бути зв'язані або функціональною, або статистичною залежністю, або ж бути взагалі незалежними.

Якщо кожному можливому значенню x величини X відповідає певне значення y іншої величини Y, то кажуть, що випадкові величини X та Y зв'язані функціональною залежністю.

Строга функціональна залежність реалізується дуже рідко, оскільки випадкові величини X та Y (або ж одна з них) зазнають впливу багатьох випадкових факторів, серед яких можуть бути і спільні. У цьому випадку між двома величинами X та Y виникає статистична залежність.

Означення. Залежність між випадковими величинами X та Y, за якою кожному значенню однієї величини відповідає розподіл іншої, називається статистичною.

Зокрема, якщо кожному можливому значенню однієї величини ставиться у відповідність середнє значення іншої, то така статистична залежність називається кореляційною.

Для випадку кореляційної залежності, якщо величина X прийняла значення x, то математичне сподівання величини Y є при цьому функцією від x:

$$M_x Y = f(x). \qquad (7.1)$$

Рівняння (7.1) називається рівнянням регресії Y на X.

Оскільки математичне сподівання є істинним (справжнім) значенням величини Y, що спостерігається, то рівняння регресії (7.1) дає справжню залежність між величинами X та Y. Тому кінцевою метою багатьох досліджень є знаходження вибіркового рівняння регресії (7.1), яке прийнято записувати у вигляді

$$\bar{y}_x = f(x). \qquad (7.2)$$

Тут $\bar{y}_x$ – умовне середнє (це середнє арифметичне значень випадкової величини Y, що відповідають значенню $X = x$); f – функція регресії Y на X.

Графік (7.2) називають вибірковою лінією регресії Y на X.

Аналогічно рівняння

$$\bar{x}_y = \varphi(y) \tag{7.3}$$

називається вибірковим рівнянням регресії X на Y. При цьому рівнянням (7.2), (7.3) у загальному випадку відповідають дві різні лінії на площині XOY.

Нехай результати вибірки подано у вигляді таблиці, що називається кореляційною таблицею.

Кореляційна таблиця.

X / Y	y_1	y_2	…	y_k	n_{x_i}	$\bar{y}_{x_i}$
x_1	n_{11}	n_{12}	…	n_{1k}	n_{x_1}	$\bar{y}_{x_1}$
x_2	n_{21}	n_{22}	…	n_{2k}	n_{x_2}	$\bar{y}_{x_2}$
…	…	…	…	…	…	…
x_m	n_{m1}	n_{m2}	…	n_{mk}	n_{x_m}	$\bar{y}_{x_m}$
n_{y_j}	n_{y_1}	n_{y_2}	…	n_{y_k}	n	–

Якщо розглядати таблицю за рядками, то кожному значенню x_i відповідає деякий розподіл випадкової величини Y. Обчислимо для цих розподілів умовні середні значення

$$\bar{y}_{x_i} = \frac{\sum_{j=1}^{k} y_j n_{ij}}{n_{x_i}},\ i = 1,2,...,m. \tag{7.4}$$

Отже, маємо залежність (7.2). Аналогічно, розглядаючи таблицю за стовпцями, визначаємо умовні середні величини

$$\bar{x}_{y_j} = \frac{\sum_{i=1}^{m} x_i n_{ij}}{n_{y_j}},\ j = 1,2,...,k. \tag{7.5}$$

Приходимо до залежності вигляду (7.3).

У кореляційному аналізі при дослідженні залежності кількісних ознак X та Y розглядають дві основні задачі:

1) знайти наближену функцію регресії, що характеризує основну тенденцію залежності Y від X (або X від Y) та належить одному з відомих типів функцій (лінійна, квадратична, показникова і т.і.);

2) оцінити силу, тісноту цієї залежності, тобто визначити ступінь розсіювання можливих значень однієї випадкової величини відносно лінії регресії, якщо одна із величин набуває певних значень.

Приклад 7.1. Здійснено спостереження двох ознак із 15 колосів пшениці − виміряли довжину кожного колосу X (см) та порахували кількість зерен Y. За результатами досліду

x_i	10	9	11	8	9	10	9	11	8	10	9	10	8	9	11
y_i	24	20	27	18	20	24	20	27	20	27	24	27	20	27	30

скласти кореляційну таблицю.

Розв'язання. Упорядкуємо ці первісні дані: в І-му стовпці запишемо в порядку зростання значення x_i: 8, 9, 10, 11, а в І-му рядку − у тому ж порядку значення y_j: 18, 20, 24, 27, 30. На перетині рядків і стовпців запишемо число повторень однакових пар (x_i, y_j) в ряду спостережень. Числа n_{ij} показують скільки повторюються парні значення x_i, y_j. Наприклад, n_{23} показує скільки разів відбулася подія: «$X = x_2, Y = y_3$».

X / Y	18	20	24	27	30	n_x
8	1	2				3
9		3	1	1		5
10			2	2		4
11				2	1	3
n_y	1	5	3	5	1	15

За кореляційною таблицею бачимо, що кожному значенню ознаки X відповідає розподіл ознаки Y і навпаки, кожному значенню Y відповідає розподіл ознаки X. Так, значенню $x_1 = 8$ відповідає розподіл

y_j	18	20
n_j	1	2

а умовне середнє $\bar{y}_{x1} = \dfrac{18 \cdot 1 + 20 \cdot 2}{3} = 19{,}3$.

Далі, при $x_2 = 9$ маємо

y_j	20	24	27
n_j	3	1	1

$$\bar{y}_{x2} = \frac{20 \cdot 3 + 24 \cdot 1 + 27 \cdot 1}{5} = 22{,}2;$$

при $x_3 = 10$

y_j	24	27
n_j	2	2

$$\bar{y}_{x_3} = \frac{24 \cdot 2 + 27 \cdot 2}{4} = 25{,}5;$$

при $x_4 = 11$

y_j	27	30
n_j	2	1

$$\bar{y}_{x_4} = \frac{27 \cdot 2 + 30 \cdot 1}{3} = 28{,}0.$$

Отже, маємо таку нову таблицю, що визначає відповідність між значеннями x_i та умовними середніми $\bar{y}_{x_i}$:

x_i	8	9	10	11
$\bar{y}_{x_i}$	19,3	22,2	25,5	28,0

За формулами (7.5) знаходимо умовні середні $\bar{x}_{y_j}$ та складаємо таблицю відповідності між значеннями y_j та $\bar{x}_{y_j}$:

y_j	18	20	24	27	30
$\bar{x}_{y_j}$	8	8,6	9,7	10,2	11,0

Заносимо умовні середні $\bar{y}_{x_i}$ та $\bar{x}_{y_j}$ в кореляційну таблицю:

X / Y	18	20	24	27	30	n_x	$\bar{y}_{x_i}$
8	1	2				3	19,3
9		3	1	1		5	22,2
10			2	2		4	25,5
11				2	1	3	28,0
n_y	1	5	3	5	1	15	
$\bar{x}_{y_j}$	8	8,6	9,7	10,2	11,0		

Зауваження. Кореляційна таблиця може бути складена як для дискретних, так і для неперервних випадкових величин. У випадку дослідження неперервних ознак X , Y від інтервального переходять до дискретного статистичного розподілу, замінюючи частинний інтервал його центром (серединою).

7.2. Знаходження вибіркового лінійного рівняння регресії.

Нехай вивчається вибірка об'єму n з двох кількісних ознак X та Y : $(x_1; y_1),(x_2; y_2),...,(x_n; y_n)$, причому значення $x_i; y_i, i = \overline{1,n}$ зустрічаються по одному разу.

У цьому випадку немає необхідності групувати дані та використовувати поняття умовної середньої. Тому шукане рівняння регресії можна записати так:

$$y = f(x) \tag{7.6}$$

або

$$x = \varphi(y). \tag{7.7}$$

Наближений вигляд згладжувальної функції f (або φ) можна підібрати, виходячи з теоретичних міркувань або ж за характером розташування на координатній площині експериментальних точок $(x_1; y_1)$, $(x_2; y_2)$, … , $(x_n; y_n)$. Це так зване поле розсіювання.

При вибраному вигляді згладжувальної функції $y = f(x,a,b)$ параметри a та b потрібно підібрати так, щоб сума квадратів відхилень y_i від $f(x_i,a,b)$ була найменшою:

$$S = S(a,b) = \sum_{i=1}^{n} [y_i - f(x_i,a,b)]^2 \to \min_{a,b} \tag{7.8}$$

У цьому розумінні функція f за методом найменших квадратів "найкращим чином" описує відповідний процес.

Задача (7.8) – це задача на безумовний екстремум функції двох змінних $S = S(a,b)$. На підставі необхідної умови екстремуму невідомі параметри a, b знаходимо за умовою

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0 , \\ \dfrac{\partial S}{\partial b} = 0 . \end{cases} \tag{7.9}$$

Нехай точки $(x_i; y_i)$, $i = 1,2,...,n$ групуються навколо прямої лінії, як на рис.7.1.

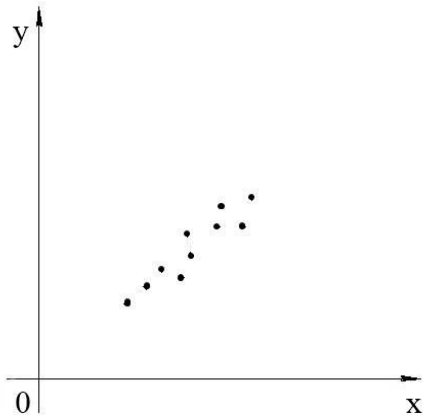

Рис. 7.1.

Тоді емпіричну функцію регресії шукають у вигляді

$$y = ax + b, \tag{7.10}$$

де a, b – невідомі параметри.

Для випадку лінійної згладжувальної функції маємо

$$S(a,b) = \sum_{i=1}^{n} \left[y_i - (ax_i + b) \right]^2,$$

а система (7.9) набуває вигляду

$$\begin{cases} \sum_{i=1}^{n} (y_i - ax_i - b) x_i = 0, \\ \sum_{i=1}^{n} (y_i - ax_i - b) = 0 \end{cases}$$

або, остаточно

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i. \end{cases} \tag{7.11}$$

Система (7.11) називається нормальною системою методу найменших квадратів для відшукання параметрів лінійної залежності. Це неоднорідна система двох лінійних рівнянь відносно невідомих a, b. Знайшовши розв'язок системи (7.11)

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \; b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}, \tag{7.12}$$

дістанемо лінійну функцію регресії Y на X.

Аналогічно знаходиться емпірична лінійна функція регресії X на Y $x = cy + d$. Для цього випадку параметри c, d є розв'язками системи

$$\begin{cases} c\sum_{i=1}^{n} y_i^2 + d\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i y_i \ , \\ c\sum_{i=1}^{n} y_i + d \cdot n = \sum_{i=1}^{n} x_i \ . \end{cases} \tag{7.13}$$

і знаходяться за формулами

$$c = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}, \ d = \frac{\sum_{i=1}^{n} x_i - c\sum_{i=1}^{n} y_i}{n}. \tag{7.14}$$

Приклад 7.2. За даними спостережень величин X та Y

x_i	1	2	3	4
y_i	3	4	6	8

знайти функцію регресії Y на X .

Розв'язання. Побудуємо точки $M_i(x_i, y_i)$, $i = 1,2,3,4$ в прямокутній системі координат на площині XOY (рис. 7.2).

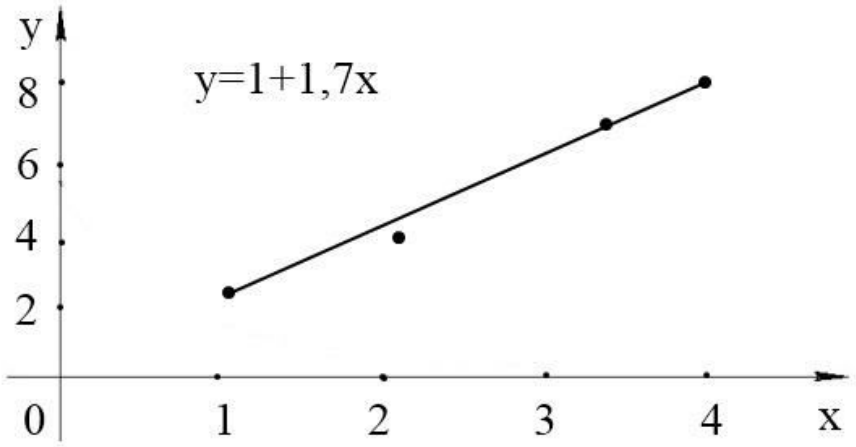

Рис. 7.2.

Ці точки групуються навколо прямої лінії, тому будемо підбирати лінійну функцію регресії . Обчислимо суми, що входять у формули (7.12). Для зручності результати обчислень заносимо в таблицю:

	x_i	y_i	$x_i\ y_i$	x_i^2
	1	3	3	1
	2	4	8	4
	3	6	18	9

	4	8	32	16
$\sum$	10	21	61	30

Знаходимо розв'язок нормальної системи за готовими формулами (7.12):

$$a=\frac{4\cdot 61-10\cdot 21}{4\cdot 30-10^2}=\frac{244-210}{120-100}=\frac{34}{20}=1{,}7\,,\ b=\frac{21-1{,}7\cdot 10}{4}=\frac{4}{4}=1\,.$$

Отже, емпірична регресія (оцінка істинної регресії) має вигляд $y=1{,}7x+1$. Графічно емпіричну регресію зображують у вигляді прямої, що групує спостережувані дані (рис. 7.2).

Якщо число спостережень n велике ($n\geq 50$), експериментальні дані прийнято групувати та записувати у вигляді кореляційної таблиці (див. п.7.1). У цій таблиці враховується частота n_x повторення значення x випадкової величини X у вибірці; частота n_y повторення значення y випадкової величини Y і частота n_{xy} повторення пари (x,y) значень випадкових величин (X,Y) у вибірці. При цьому $\sum n_x=\sum n_y=\sum n_{xy}=n$ (для спрощення поточні індекси в сумах опускаємо).

У випадку згрупованих даних система для знаходження коефіцієнтів a, b лінійної емпіричної функції регресії Y на X

$$\bar{y}_x=ax+b \tag{7.15}$$

набуває вигляду

$$\begin{cases} a\sum n_x x^2+b\sum n_x x=\sum n_{xy}xy\,, \\ a\sum n_x x+nb=\sum n_y y\,, \end{cases} \tag{7.16}$$

її розв'язок:

$$a=\frac{n\sum n_{xy}xy-\sum n_x x\sum n_y y}{n\sum n_x x^2-\left(\sum n_x x\right)^2}\,;\ b=\frac{1}{n}\left(\sum n_y y-a\sum n_x x\right). \tag{7.17}$$

Аналогічно трансформується система (7.13) для знаходження коефіцієнтів c, d лінійної регресії X на Y $\bar{x}_y=cy+d$ та формули (7.14) для обчислення цих коефіцієнтів.

Слід зазначити, що $\bar{y}_x=ax+b$ та $\bar{x}_y=cy+d$ – різні прямі (рис.7.3).

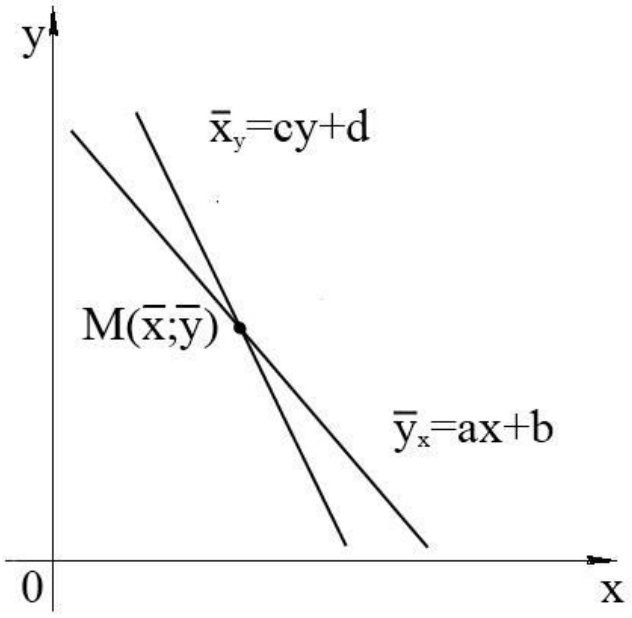

Рис. 7.3.

Перша пряма одержується в результаті розв'язання задачі про мінімізацію суми квадратів відхилень по вертикалі, а друга − по горизонталі. Прямі лінійної регресії перетинаються в точці $M(\bar{x},\bar{y})$, що називається центром кореляції.

7.3. Вибірковий коефіцієнт кореляції

Основними характеристиками, що описують силу зв'язку між випадковими величинами X та Y, є кореляційний момент

$$\mu_{xy} = M[(X - M(X))(Y - M(Y))] \tag{7.18}$$

і коефіцієнт кореляції

$$r_{xy} = \frac{\mu_{xy}}{\sigma(X)\sigma(Y)}, \tag{7.19}$$

для обчислення яких потрібно знати закони розподілу величин X та Y.

При обробці експериментальних даних, як правило, закони розподілу невідомі. Тому для оцінки сили зв'язку між величинами X та Y застосовують точкові оцінки μ_{xy} і r_{xy} − вибірковий кореляційний момент

$$K_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{7.20}$$

та вибірковий коефіцієнт кореляції

$$r_B = \frac{K_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y\right)^2}} \tag{7.21}$$

або

$$r_B = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - y\right)^2}}, \tag{7.22}$$

де $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – відповідні середні значення (вибіркові середні).

Вибірковий коефіцієнт кореляції, як оцінка коефіцієнта кореляції генеральної сукупності, слугує для вимірювання лінійного зв'язку між величинами – кількісними ознаками X та Y.

Властивості вибіркового коефіцієнта кореляції:

1. $-1 \le r_B \le 1$ (або $|r_B| \le 1$).

2. Чим більшою є величина r_B, тим тісніший зв'язок між досліджуваними ознаками X та Y.

3. Якщо $|r_B| = 1$, то кореляційна залежність між X та Y стає лінійною функціональною.

4. Якщо $r_B = 0$, то між досліджуваними ознаками X та Y немає лінійної кореляційної залежності, але умова $r_B = 0$ не виключає існування будь-якої іншої кореляційної залежності (параболічної, показникової і т.і.).

Якщо з деяких теоретичних міркувань заздалегідь відомо, що величини X та Y мають нормальний розподіл, то рівність $r_B = 0$ свідчить про відсутність будь-якої залежності між ознаками X та Y (тобто величини X, Y – незалежні).

Щоб одержати ще одну формулу для розрахунку r_B, скористаємося формулами

$$D(X) = M\left(X^2\right) - \left[M(X)\right]^2,\ \sigma(X) = \sqrt{D(X)},\ D(Y) = M\left(Y^2\right) - \left[M(Y)\right]^2,\ \sigma(Y) = \sqrt{D(Y)}.$$

та перетворимо (7.18) до вигляду

$$\mu_{xy} = M(X \cdot Y) - M(X) \cdot M(Y).$$

Тоді

$$\sigma_x^2 = \overline{x^2} - \left(\bar{x}\right)^2,\ \sigma_y^2 = \overline{y^2} - \left(\bar{y}\right)^2,\ K_{xy} = = \overline{x \cdot y} - \bar{x} \cdot \bar{y},$$

а формула для обчислення вибіркового коефіцієнта кореляції буде такою:

$$r_B = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}} \tag{7.23}$$

або

$$r_B = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}, \tag{7.24}$$

де $\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}$, $\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}$ – середні квадратичні відхилення величин X та Y відповідно.

Щоб встановити зв'язок r_B з вибірковим лінійним рівнянням регресії $\bar{y}_x = ax + b$, запишемо нормальну систему для визначення коефіцієнтів a, b у вигляді

$$\begin{cases} a\overline{x^2} + b\bar{x} = \overline{x \cdot y}, \\ a\bar{x} + b = \bar{y}. \end{cases} \tag{7.25}$$

Далі знайдемо з другого рівняння системи $b = \bar{y} - a\bar{x}$ та підставимо в рівняння регресії:

$$\bar{y}_x = ax + \bar{y} - a\bar{x}$$

або

$$\bar{y}_x - \bar{y} = a(x - \bar{x}). \tag{7.26}$$

Для визначення a помножимо друге рівняння системи (7.25) на $\bar{x}$ та віднімемо його від першого рівняння:

$$a\overline{x^2} - a(\bar{x})^2 = \overline{x \cdot y} - \bar{x} \cdot \bar{y}.$$

Звідси

$$a = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

або

$$a = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sigma_x^2}. \tag{7.27}$$

Порівнюючи (7.27) з виразом для коефіцієнта кореляції (7.24), приходимо до такої формули-звязку:

$$a = \frac{\sigma_y}{\sigma_x} r_{xy}. \tag{7.28}$$

Тепер вираз для a (7.28) підставимо в лінійне рівняння регресії Y на X:

$$\bar{y}_x - \bar{y} = r_{xy} \frac{\sigma_y}{\sigma_x}(x - \bar{x}). \tag{7.29}$$

Аналогічно можна знайти вибіркове рівняння прямої лінії регресії X на Y:

$$\bar{x}_y - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right). \qquad (7.30)$$

Приклад 7.3. Знання 10 студентів перевірені за двома тестами A та B. Оцінки за стобальною системою виявились такими:

№ студента	1	2	3	4	5	6	7	8	9	10
кількість балів за тестом A, x_i	60	58	57	55	56	58	55	57	55	59
кількість балів за тестом B, y_i	56	53	54	51	54	59	55	55	56	57

Обчислити коефіцієнт кореляції.

Розв'язання. Розрахунок коефіцієнта кореляції будемо виконувати за формулою (7.22).

Для визначення величин, що входять до цієї формули, складемо допоміжну таблицю, в якій результати спостереження x_i та y_i запишемо стовпцями. В кінці кожного стовпця обчислимо суми для розрахунку середніх $\bar{x}$ та $\bar{y}$.

№ спостереження	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	60	56	3	9	1	1	3
2	58	53	1	1	-2	4	-2
3	57	54	0	0	-1	1	0
4	55	51	-2	4	-4	16	8
5	56	54	-1	1	-1	1	1
6	58	59	1	1	4	16	4
7	55	55	-2	4	0	0	0
8	57	55	0	0	0	0	0

9	55	56	-2	4	1	1	-2
10	59	57	2	4	2	4	4
$\sum$	570	550		28		44	16

Праворуч знаходяться стовпці, в яких обчислюються різниці $(x_i - \bar{x})$ та $(y_i - \bar{y})$, їх квадрати та добутки. Щоб одержати величини, необхідні для розрахунку коефіцієнта кореляції, складемо значення відповідних стовпців.

Визначаючи середні $\bar{x} = \frac{570}{10} = 57, \quad \bar{y} = \frac{550}{10} = 55$, за допоміжною таблицею дістанемо

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 28; \quad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 44; \quad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 16.$$

Підставивши ці значення у формулу для обчислення коефіцієнта кореляції (7.22), одержимо

$$r = \frac{16}{\sqrt{28}\sqrt{44}} = 0{,}457 \neq 0.$$

Для встановлення зв'язку між випадковими величинами X та Y скористуємося критерієм

$$|r|\sqrt{n-1} \geq 3.$$

Дістанемо: $|r|\sqrt{n-1} = 0{,}457\sqrt{9} = 1{,}371 < 3.$ Отже, зв'язок між величинами X та Y недостатньо ймовірний.

Практичне заняття №8. Кореляційний та регресійний аналіз

Опитування з теорії.

1. Яку залежність між випадковими величинами називають функціональною? статистичною? кореляційною?
2. Як означається рівняння регресії Y на X (X на Y)?
3. Що називають умовною середньою?
4. Як записують вибіркове рівняння регресії?
5. Як називається графік вибіркового рівняння регресії?
6. Опишіть структуру кореляційної таблиці. Чи можливо скласти кореляційну таблицю для неперервних випадкових величин?
7. Які дві основні задачі розглядають у кореляційному аналізі?
8. Як знайти наближений вигляд згладжувальної функції? Що таке поле розсіювання?
9. У чому полягає суть методу найменших квадратів?
10. Який вигляд має нормальна система методу найменших квадратів для випадку лінійної згладжувальної функції?
11. Що розуміють під прямолінійною та криволінійною кореляціями?
12. Як означається вибірковий коефіцієнт кореляції? Оцінкою чого слугує вибірковий коефіцієнт кореляції?
13. Сформулювати властивості вибіркового коефіцієнта кореляції.
14. Як означається кореляційне відношення випадкової величини Y до випадкової величини X (випадкової величини X до випадкової величини Y)?
15. . Сформулювати основні властивості кореляційних відношень.

Завдання для аудиторної роботи.

Приклад 7.1. За даними 24 спостережень ознак X та Y:

X	0,5	0,7	0,6	0,7	0,8	0,8	0,6	0,5	0,8	0,7	0,6	0,7
Y	0,8	0,5	0,6	0,7	0,5	0,6	0,7	0,8	0,5	0,6	0,8	0,7

X	0,5	0,8	0,7	0,9	0,5	0,9	0,6	0,9	0,6	0,6	0,7	0,7
Y	0,6	0,6	0,5	0,5	0,6	0,5	0,7	0,7	0,8	0,8	0,8	0,7

скласти кореляційну таблицю.

Розв'язання. Упорядкуємо ці первісні дані: в І-му стовпці запишемо в порядку зростання значення ознаки $X - x_i$: 0,5; 0,6; 0,7; 0,8; 0,9, а в І-му рядку – у тому ж порядку значення ознаки $Y - y_j$: 0,5; 0,6; 0,7; 0,8. На перетині рядків і стовпців запишемо число повторень однакових пар (x_i, y_j) в ряду спостережень. Числа n_{ij} показують скільки повторюються парні

значення x_i , y_j . Наприклад, n_{23} показує скільки разів відбулася подія: «$X = x_2, Y = y_3$». Для нашого прикладу $n_{23} = 2$, а $X = 0{,}6$, $Y = 0{,}7$.

X / Y	0,5	0,6	0,7	0,8	n_x
0,5	0	2	0	2	4
0,6	0	1	2	3	6
0,7	2	1	3	1	7
0,8	2	2	0	0	4
0,9	2	0	1	0	3
n_y	6	6	6	6	24

Приклад 7.2. За даними 79 спостережень ознак $X = \sigma_S / \sigma_B$ та Y:

X / Y	0,5	0,6	0,7	0,8	n_x
0,5	0	2	0	8	10
0,6	0	4	2	9	15
0,7	2	12	3	1	18
0,8	21	14	0	0	35
0,9	1	0	0	0	1
n_y	24	32	5	18	79

(σ_S − межа текучості сталі, σ_B − межа міцності сталі, Y − відсотковий вміст вуглецю у сталі) знайти коефіцієнт кореляції.

Розв'язання. Застосовуючи дані таблиці, знаходимо вибіркові середні (за формулою для згрупованих даних $\bar{x} = \frac{1}{n}\sum_{i=1}^{k_1} x_i n_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{k_2} y_i n_i$):

$$\bar{x} = \frac{1}{n}\sum n_x x = \frac{0{,}5 \cdot 10 + 0{,}6 \cdot 15 + 0{,}7 \cdot 18 + 0{,}8 \cdot 35 + 0{,}9 \cdot 1}{79} = \frac{55{,}5}{79} \approx 0{,}703,$$

$$\bar{y} = \frac{1}{n}\sum n_y y = \frac{0{,}5 \cdot 24 + 0{,}6 \cdot 32 + 0{,}7 \cdot 5 + 0{,}8 \cdot 18}{79} = \frac{49{,}1}{79} \approx 0{,}622.$$

Визначимо допоміжні характеристики:

$$\frac{1}{n}\sum n_x x^2 = \frac{0{,}5^2 \cdot 10 + 0{,}6^2 \cdot 15 + 0{,}7^2 \cdot 18 + 0{,}8^2 \cdot 35 + 0{,}9^2 \cdot 1}{79} = \frac{39{,}93}{79} \approx 0{,}505;$$

$$\frac{1}{n}\sum n_y y^2 = \frac{0{,}5^2 \cdot 24 + 0{,}6^2 \cdot 32 + 0{,}7^2 \cdot 5 + 0{,}8^2 \cdot 18}{79} = \frac{31{,}49}{79} \approx 0{,}398;$$

$$\frac{1}{n}\sum n_{xy} xy = \frac{1}{79}(0{,}5 \cdot 0{,}6 \cdot 2 + 0{,}5 \cdot 0{,}8 \cdot 8 + 0{,}6 \cdot 0{,}6 \cdot 4 + 0{,}6 \cdot 0{,}7 \cdot 2 + 0{,}6 \cdot 0{,}8 \cdot 9 + + 0{,}7 \cdot 0{,}5 \cdot 2 +$$

$$+ 0{,}7 \cdot 0{,}6 \cdot 12 + 0{,}7 \cdot 0{,}7 \cdot 3 + 0{,}7 \cdot 0{,}8 \cdot 1 + 0{,}8 \cdot 0{,}5 \cdot 21 + 0{,}8 \cdot 0{,}6 \cdot 14 + 0{,}9 \cdot 0{,}5 \cdot 1) =$$

$$= \frac{1}{79} \cdot 33{,}72 \approx 0{,}427.$$

Знаходимо дисперсії і вибірковий кореляційний момент:

$$\sigma_x^2 = \overline{x^2} - (\bar{x})^2 = 0{,}505 - (0{,}703)^2 = 0{,}505 - 0{,}493 = 0{,}012,\ \sigma_x = 0{,}11;$$

$$\sigma_y^2 = \overline{y^2} - (\bar{y})^2 = 0{,}398 - (0{,}622)^2 = 0{,}398 - 0{,}387 = 0{,}011,\ \sigma_y = 0{,}105;$$

$$K_{xy} = = \overline{x \cdot y} - \bar{x} \cdot \bar{y} = 0{,}427 - 0{,}703 \cdot 0{,}622 = 0{,}427 - 0{,}437 = -0{,}01.$$

Вибірковий коефіцієнт кореляції визначаємо за формулою:

$$r_B = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}},$$

тобто

$$r_B = -\frac{0{,}01}{0{,}11 \cdot 0{,}105} = -0{,}867.$$

Обчислимо значення добутку

$$|r_B| \cdot \sqrt{n-1} = 0{,}867 \cdot \sqrt{78} = 0{,}867 \cdot 8{,}84 = 7{,}66.$$

Оскільки $|r_B| \cdot \sqrt{n-1} > 3$, то зв'язок між величинами X та Y достатньо ймовірний.

Приклад 7.3. Вимірювання електричного опору R провідника при певних температурах t привело до таких результатів :

t , $^\circ C$	19	25	30	36	40
R , *ом*	76,3	77,8	79,8	80,8	82,3

Потрібно підібрати лінійну емпіричну формулу та знайти її параметри за методом найменших квадратів.

Розв'язання. Маємо випадок не згрупованих даних. Лінійна функція $R = at + b$ слугуватиме емпіричною формулою залежності опору R від температури t.

Згідно з методом найменших квадратів параметри a та b цієї формули є розв'язком системи рівнянь вигляду:

$$\sum_{i=1}^{5} R_i t_i = a\sum_{i=1}^{5} t_i^2 + b\sum_{i=1}^{5} t_i,$$
$$\sum_{i=1}^{5} R_i = a\sum_{i=1}^{5} t_i + 5b,$$

що аналітично записується так :

$$a = \frac{5\sum_{i=1}^{5} R_i t_i - \sum_{i=1}^{5} R_i \cdot \sum_{i=1}^{5} t_i}{5\sum_{i=1}^{5} t_i^2 - (\sum_{i=1}^{5} t_i)^2}, \qquad b = \frac{1}{5}(\sum_{i=1}^{5} R_i - a\sum_{i=1}^{5} t_i).$$

Для обчислення сум в наведених формулах складено таку допоміжну таблицю:

i	t_i	R_i	t_i^2	$R_i\ t_i$
1	19	76,3	361	1449,7
2	25	77,8	625	1945
3	30	79,8	900	2394
4	36	80,8	1296	2908,8
5	40	82,3	1600	3292
$\sum$	150	397	4782	11989,5

Підставивши обчислені суми в формули для коефіцієнтів a і b, дістанемо

$$a = \frac{5 \cdot 11989{,}5 - 397 \cdot 150}{5 \cdot 4782 - (150)^2} = \frac{397{,}5}{1410} = 0{,}282,$$

$$b = \frac{1}{5}(397 - 0{,}282 \cdot 150) = 70{,}94$$

Отже , шукана емпірична формула матиме такий вигляд:

$$R = 0{,}282 \cdot t + 70{,}94.$$

Приклад 7.4. За даними прикладу 7.2 знайти лінійні рівняння регресії.

Розв'язання. Лінійне рівняння регресії Y на X має вигляд (7.29):

$$\bar{y}_x - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right),$$

тобто

$$\bar{y}_x - 0{,}622 = -0{,}867 \cdot \frac{0{,}105}{0{,}11}\left(x - 0{,}703\right),$$

$$\bar{y}_x = -0{,}828x + 1{,}204\,.$$

Вибіркове рівняння прямої лінії регресії X на Y:

$$\bar{x}_y - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right),$$

тобто

$$\bar{x}_y - 0{,}703 = -0{,}867 \cdot \frac{0{,}11}{0{,}105}\left(y - 0{,}622\right),$$

$$\bar{x}_y = -0{,}908\,y + 1{,}268\,.$$

Приклад 7.5. Знайти вибіркове рівняння регресії $\bar{y}_x = ax^2 + bx + c$ за даними кореляційної таблиці:

Y/X	2	3	5	n_y
25	20	0	0	20
45	0	30	1	31
110	0	1	48	49
n_x	20	31	49	$n = 100$

Розв'язання. Оскільки дані згруповані, система для знаходження невідомих параметрів a, b, c набуває вигляду:

$$\begin{cases} a\sum n_x x^4 + b\sum n_x x^3 + c\sum n_x x^2 = \sum n_x \bar{y}_x x^2, \\ a\sum n_x x^3 + b\sum n_x x^2 + c\sum n_x x = \sum n_x \bar{y}_x x, \\ a\sum n_x x^2 + b\sum n_x x + nc = \sum n_x \bar{y}_x. \end{cases}$$

Складемо розрахункову таблицю

x	n_x	$\bar{y}_x$	$n_x\ x$	$n_x\ x^2$	$n_x\ x^3$	$n_x\ x^4$	$n_x\ \bar{y}_x$	$n_x\ \bar{y}_x\ x$	$n_x\ \bar{y}_x\ x^2$
2	20	25	40	80	160	320	500	1000	2000
3	31	47,1	93	279	837	2511	1460	4380	13141
5	49	108,67	245	1225	6125	30625	5325	26624	133121
$\sum$	100		378	1584	7122	33456	7285	32004	148262

Тут $\bar{y}_x$ – умовні середні: $\bar{y}_{x=2} = \frac{25 \cdot 20}{20} = 25$; $\bar{y}_{x=3} = \frac{45 \cdot 30 + 110 \cdot 1}{31} = \frac{1460}{31} \approx 47{,}1$;

$\bar{y}_{x=5} = \frac{45 \cdot 1 + 110 \cdot 48}{49} = \frac{5325}{49} \approx 108{,}67$.

Підставивши числа останнього рядка таблиці, дістанемо систему рівнянь

$$\begin{cases} 33456a + 7122b + 1584c = 148262, \\ 7122a + 1584b + 378c = 32004, \\ 1584a + 378b + 100c = 7285, \end{cases}$$

розв'язки якої знаходимо, наприклад, за методом Гаусса: $a = 2{,}94$; $b = 7{,}27$; $c = -1{,}25$.

Отже, маємо шукане рівняння квадратичної регресії вигляду:

$$\bar{y}_x = 2{,}94x^2 + 7{,}27x = 1{,}25.$$

Завдання для самостійної роботи.

1. Результати лабораторних аналізів десяти зразків речовини щодо вмісту компонент *X* та *Y* (у відсотках) зведені в таблицю

X	4	6	3	10	6	2	3	6	7	3
Y	11	17	10	24	16	6	7	21	20	8

Вважаючи залежність між випадковими величинами *X* та *Y* близькою до лінійної, знайти а) лінійні рівняння регресії *Y* на *X* та *X* на *Y*; б) обчислити вибірковий коефіцієнт кореляції.
Відповідь: а) $y = 2{,}5x + 1{,}5$; $x = 0{,}3629y + 0{,}0806$; б) $r_B \approx 0{,}95$.

2. Знайти лінійні рівняння регресії *Y* на *X* та *X* на *Y* за вибірковими даними:

а)

X	2,7	4,6	6,3	7,8	9,2	10,6	12,0	13,4	14,7
Y	17,0	16,2	13,3	13,0	9,7	9,9	6,2	5,8	5,7

б)

X	7,9	11,6	12,8	14,9	16,3	18,6	20,3	21,9	23,6
Y	13,0	22,8	24,8	28,6	31,6	38,7	40,0	44,9	43,0

Відповідь: а) $y = -1{,}06x + 20{,}3$; $x = -0{,}971y + 19{,}477$;

б) $y = 2{,}03x - 1{,}49$; $x = 0{,}478y + 1{,}17$.

3. Обчислити вибірковий коефіцієнт кореляції та скласти лінійні рівняння регресії за даними вибірки:

а)

X	67	54	72	64	39	22	58	43	46	34
Y	24	15	23	19	16	11	20	16	17	13

б)

X	60	58	57	55	56	58	55	57	55	59
Y	56	53	54	51	54	59	55	55	56	57

в)

Y \ X	2	3	4	5
2				
3		6	2	
4		6	23	5
5			3	17

г)

Y \ X	0,5	0,6	0,7	0,8	0,9
0,5			2	21	1
0,6	2	4	12	14	
0,7		2	3		
0,8	8	9	1		

д)

Інтервали зміни X / Інтервали зміни Y	1 – 2	2 – 3	3 – 4	4 – 5	5 – 6	6 – 7	7 – 8	8 – 9
10 – 20	4	5						
20 – 30	1	3	1					
30 – 40	2	3	6	5	3	1		
40 – 50		5	9	19	8	7	2	1
50 – 60		1	2	7	16	9	4	2
60 – 70			1	5	6	4	2	2
70 – 80							1	3

є)

Інтервали зміни X / Інтервали зміни Y	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
21 – 29					2
29 – 37			8	8	
37 – 45		6	13	3	
45 - 53	2	3	4	1	

Відповідь: а) $r_B = 0{,}92$; $y = 0{,}24x + 5{,}42$; $x = 3{,}53y - 11{,}56$;

б) $r_B = 0{,}455$; $y = 0{,}57x + 22{,}51$; $x = 0{,}36y + 37{,}2$;

в) $r_B = 0{,}75$; $y = 0{,}6x + 1{,}73$; $x = 0{,}8y + 0{,}79$;

г) $r_B = -0{,}83$; $y = -0{,}83x + 1{,}2$; $x = -0{,}77y + 1{,}18$;

д) $r_B = 0{,}67$; $y = 5{,}16x + 22{,}12$; $x = 0{,}09y + 0{,}63$;

є) $r_B = -0{,}62$; $y = -0{,}45x + 50{,}48$; $x = -0{,}84y + 58{,}33$.

4. Знайти вибіркову функцію регресії $\bar{y}_x = ax^2 + bx + c$ та вибіркове кореляційне відношення η^*_{yx} за даними кореляційної таблиці:

а)

X/Y	2	3	5
25	20		
45		30	1
110		31	48

б)

X/Y	0	4	5
1	50	5	1
35		44	
50		5	45

Відповідь: а) $\bar{y}_x = 2{,}94x^2 + 7{,}27x - 1{,}25$; $\eta^*_{yx} = 0{,}89$;

б) $\bar{y}_x = 1{,}53x^2 + 1{,}95x + 1$; $\eta^*_{yx} = 0{,}86$.

5. Знайти вибіркову функцію регресії $\bar{x}_y = ay^2 + by + c$ та вибіркове кореляційне відношення η^*_{xy} за даними кореляційної таблиці:

а)

X/Y	6	30	50
1	15		
3	1	14	
4		2	18

б)

X/Y	1	9	19
0	13		
2	2	10	
3	1	1	23

Відповідь: а) $\bar{x}_y = 2{,}8y^2 + 0{,}02y + 3{,}18$; $\eta^*_{xy} = 0{,}96$;

б) $\bar{x}_y = 2{,}29y^2 - 1{,}25y + 1$; $\eta^*_{xy} = 0{,}92$.

Список джерел.

1. Сулима І.М., Яковенко В.М. Вища математика. Теорія ймовірностей. Математична статистика. Навчальний посібник. К.: Вид. Центр НАУ, 2004. – 238 с.

2. Сулима І.М., Ковтун І.І., Нікітіна І.А., Скороход Т.А., Яковенко В.М. Прикладна математика. Теорія ймовірностей. Математична статистика. Навчально-методичний посібник. К.: Вид. Центр НАУ, 2005. – 148 с.

3. Валєєв К. Г., Джалладова І. А. Вища математика: Навч. посібник: У 2-х ч. — Ч. 2. — К.: КНЕУ, 2002. — 451 с
4. Сулима І.М., Панталієнко Л.А., Яковенко В.М. Методичні рекомендації та індивідуальні завдання з дисципліни „Прикладна математика” для студентів інженерних факультетів. – К.: Вид. центр НАУ, 2001. – 67 с.
5. Сулима І.М., Панталієнко Л.А., Якимів Р.Я. Методичні рекомендації та контрольні завдання з дисципліни „Прикладна математика” для студентів факультету електрифікації та автоматизації сільськогосподарського виробництва заочної форми навчання. – К.: Вид. центр НАУ, 2003. – 62 с.
6. Астахов В.М. Теорія ймовірностей і математична статистика. Навчально-методичний посібник / В.М.Астахов, Г.С. Буланов, В.О. Паламарчук – Краматорськ: ДДМА, 2009. – 64 с.

http://www.dgma.donetsk.ua/metod/vm/tims.pdf

Викладач: доц. Панталієнко Л.А.
Прошу надсилати виконані завдання за адресою: wnyrk15@gmail.com