

Лекція 3

Емпіричні та вирівнюючі (теоретичні) частоти

1. Дискретна випадкова величина

Розглянемо дискретну випадкову величину X , закон розподілу якої невідомий. Нехай виконано n випробувань, в яких величина X прийняла n_1 раз значення x_1 , n_2 рази – значення x_2 і т. д., причому $n = \sum_i n_i$.

Емпіричними частотами називають фактично спостережувані частоти.

Припустимо, що у нас є підстави припустити, що досліджувана величина розподілена за деяким певним законом. Для того, щоб перевірити, чи узгоджується це припущення з даними спостережень, обчислюють частоти спостережуваних значень, тобто знаходять теоретично скільки разів величина повинна була прийняти кожне з спостережуваних значень, якщо вона розподілена по спостережуваному закону.

Вирівнюючими (теоретичними), на відміну від фактично спостережуваних емпіричних частот, називають частоти, знайдені теоретично (обчисленнями). Їх знаходять за співвідношенням:

$$n'_i = nP_i,$$

де n – кількість випробувань, P_i – ймовірність спостереження величини x_i .

2. Неперервний розподіл.

У разі неперервного розподілу, ймовірності окремих можливих значень дорівнюють нулю. Тому весь інтервал можливих значень ділять на k інтервалів що не перетинаються, і обчислюють ймовірності попадання X в i -й частковий інтервал, а потім, як і для дискретного розподілу, множать число випробувань на ці ймовірності.

Отже, вирівнюють частоти неперервного розподілу знаходять за співвідношенням: $n'_i = nP_i$.

Зокрема, якщо є підстави припустити, що випадкова величина (генеральна сукупність) розподілена нормально, то вирівнюють частоти можуть бути знайдені за формулою: $n'_i = \frac{nh}{\sigma_b} \varphi(u_i)$, де n – кількість випробувань, h – довжина часткового

інтервалу, σ_b – вибіркове середньо квадратичне відхилення, $\varphi(u_i)$ – диференціальна функція нормованого нормального розподілу, де $u_i = \frac{x_i - \bar{x}}{\sigma_b}$

$$\varphi(u_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Побудова нормальної кривої за експериментальними даними

Один із способів побудови нормальної кривої за даними спостережень полягає в наступному:

1) знаходять $\bar{x}$ і σ_b ;

2) знаходять ординати y_i (вирівнюючі частоти) теоретичної кривої за формулою $n'_i = \frac{nh}{\sigma_b} \phi(u_i)$.

3) будують точки (x_i, y_i) в прямокутній системі координат і з'єднують їх плавною кривою.

Близькість вирівнюючих частот до спостережуваних підтверджує правильність припущення про те, що ознака розподілена нормально.

Приклад. Побудувати нормальну криву з даного розподілу:
 варіанти . . . x_i 15 20 25 30 35 40 45 50 55
 частоти . . . n_i 6 13 38 74 106 85 30 10 4

Розв'язок. Знайдемо $\bar{x}_s = 34,7$, $\sigma_b = 7,38$.

Обчислимо вирівнюючі частоти

x_i	n_i	$x_i - \bar{x}_s$	$u_i = (x_i - \bar{x}_s) / \sigma_b$	$\phi(u_i)$	$y_i = \frac{nh}{\sigma_b} \phi(u_i) = 248 * \phi(u_i)$
15	6	-19,7	-2,67	0,0113	3
20	13	-14,7	-1,99	0,0551	14
25	38	-9,7	-1,31	0,1691	42
30	74	-4,7	-0,63	0,3271	82
35	106	0,3	0,05	0,3984	99
40	85	5,3	0,73	0,3056	76
45	30	10,3	1,41	0,1476	37
50	10	15,3	2,09	0,0449	12
55	4	2,77	2,77	0,0086	2
	$n=366$				$\sum y_i = 366$

Нанесемо точки (x_i, y_i) на графік і плавно з'єднаємо.

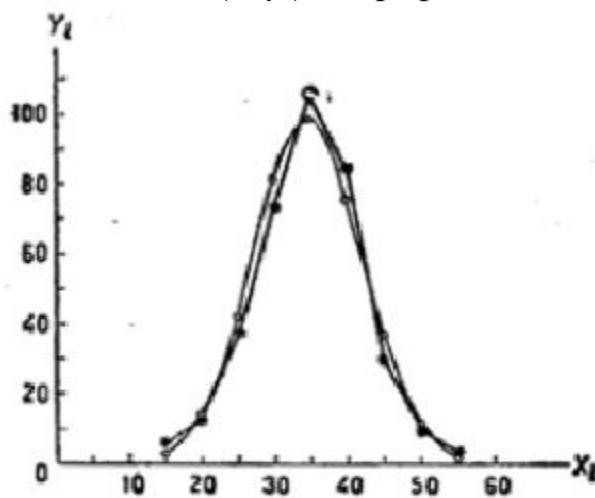


Рис. 22

В якості оцінок відхилення закону розподілу від нормального використовують асиметрію та ексцес.

Елементи теорії кореляції

Функціональна, статистична і кореляційна залежності

У багатьох задачах потрібно встановити та оцінити залежність досліджуваної випадкової величини Y від однієї або декількох інших величин. Розглянемо спочатку залежність Y від однієї випадкової (або не випадковою) величини X , а потім від декількох величин.

Дві випадкові величини можуть бути пов'язані або функціональною залежністю, або залежністю іншого роду, статистичною, або бути незалежними.

Строга функціональна залежність реалізується рідко, так як обидві величини або одна з них схильні до дії випадкових чинників, причому серед них можуть бути і загальні для обох величин (під «загальними» тут маються на увазі такі фактори, які впливають і на Y і на X). В цьому випадку виникає статистична залежність.

Наприклад, якщо Y залежить від випадкових факторів Z_1, Z_2, V_1, V_2 , а X залежить від випадкових факторів Z_1, Z_2, U_1 , то між Y і X є статистична залежність, так як серед випадкових факторів є загальні, а саме: Z_1, Z_2 .

Статистичною називають залежність, при якій зміна однієї з величин тягне за собою зміну розподілу іншої. Зокрема, статистична залежність проявляється в тому, що при зміні однієї з величин змінюється середнє значення іншої; в цьому випадку статистичну залежність називають *кореляційною*.

Наведемо приклад випадкової величини Y , яка не пов'язана з величиною X функціонально, а пов'язана кореляційно. Нехай Y -урожай зерна, X - кількість добрив. З однакових за площею ділянок землі при рівних кількостях внесених добрив знімають різний урожай, іншими словами. Y не є функцією від X . Це пояснюється впливом випадкових факторів (опаді, температура повітря та ін.). Разом з тим, як показує досвід, середній урожай є функцією від кількості добрив, отже. Y пов'язаний з X кореляційною залежністю.

Вибіркові рівняння регресії

Раніше були введені рівняння регресії Y на X і X на Y :

$$M(Y|x) = f(x), \quad M(X|y) = \varphi(y).$$

Умовне математичне сподівання $M(Y|x)$ є функцією від x , отже, його оцінка є. умовне середнє, також функція від x ; позначивши цю функцію через $f^*(x)$, отримаємо рівняння

$$\bar{y}_x = f^*(x).$$

Це рівняння називають вибіркоvim рівнянням регресії Y на X ; функцію $f^*(x)$ називають вибірковою регресією Y на X , а її графік-вибірковою лінією регресії Y на X . Аналогічно рівняння

$$\bar{x}_y = \varphi^*(y)$$

називають вибіркоvim рівнянням регресії X на Y ; функцію $\varphi^*(y)$ називають вибірковою регресією X на Y , а її графік-вибірковою лінією регресії X на Y .

Виникають наступні питання. Як знайти за даними спостережень параметри функцій $f^*(x)$ і $\varphi^*(y)$, якщо вид їх відомий? Як оцінити силу (тісноту) зв'язку між величинами X і Y і встановити, чи корельовані ці величини?

Відшукування параметрів вибіркового рівняння прямої лінії середньоквадратичної регресії по її згрупованим даним

Нехай вивчається система кількісних ознак (X, Y) . В результаті n незалежних дослідів отримані пари чисел $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Знайдемо за даними спостережень вибіркове рівняння прямої лінії середньоквадратичної регресії. Для визначеності будемо шукати рівняння

$$\bar{y}_x = kx + b$$

регресії Y на X .

Оскільки різні значення x ознаки X і відповідні їм значення y ознаки Y спостерігалися по одному разу, то групувати дані немає необхідності. Також немає потреби використовувати поняття умовної середньої, тому дані рівняння можна записати так:

$$Y = kx + b.$$

Кутовий коефіцієнт прямої лінії регресії Y на X називають вибірковим коефіцієнтом регресії Y на X і позначають через ρ_{yx} ; він є оцінкою коефіцієнта регресії.

Отже, будемо шукати вибіркове рівняння прямої лінії регресії Y на X виду

$$Y = \rho_{yx}x + b.$$

Підберемо параметри ρ_{yx} і b так, щоб сума квадратів відхилень була мінімальною (в цьому полягає сутність методу найменших квадратів). Так як кожне відхилення залежить від цих параметрів, то і сума квадратів відхилень є функція F цих параметрів (замість ρ_{yx} писатимемо ρ):

$$F(\rho, b) = \sum_{i=1}^n (Y_i - y_i)^2, \text{ або } F(\rho, b) = \sum_{i=1}^n (\rho x_i + b - y_i)^2.$$

Для відшукування мінімуму прирівняємо нулю відповідні приватні похідні:

$$\frac{dF}{d\rho} = 2 \sum_{i=1}^n (\rho x_i + b - y_i) x_i = 0;$$

$$\frac{dF}{db} = 2 \sum_{i=1}^n (\rho x_i + b - y_i) = 0.$$

Виконавши елементарні перетворення, отримаємо систему двох лінійних рівнянь щодо ρ і b

$$(\sum x^2)\rho + (\sum x)b = \sum xy; (\sum x)\rho + nb = \sum y.$$

Вирішивши цю систему, знайдемо шукані параметри:

$$\rho_{yx} = (n \cdot \sum xy - \sum x \cdot \sum y) / (n \sum x^2 - (\sum x)^2);$$

$$b = (\sum x^2 \cdot \sum y - \sum x \cdot \sum xy) / (n \sum x^2 - (\sum x)^2)$$

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

Приклад. Знайти вибіркове рівняння прямої лінії регресії Y на X за даними

$n = 5$ спостережень:

x 1,00 1,50 3,00 4,50 5,00

y 1,25 1,40 1,50 1,75 2,25

Розв'язок. Складемо розрахункову табл. 11.

x_i	y_i	x_i^2	$x_i y_i$
1,00	1,25	1,00	1,250
1,50	1,40	2,25	2,100
3,00	1,50	9,00	4,500
4,50	1,75	20,25	7,875
5,00	2,25	25,00	11,250
$\sum x_i = 15$	$\sum y_i = 8,15$	$\sum x_i^2 = 57,50$	$\sum x_i y_i = 26,975$

Знайдемо шукані параметри, для чого підставимо обчислені по таблиці суми в співвідношення

$$\rho_{xy} = (5 \cdot 26,975 - 15 \cdot 8,15) / (5 \cdot 57,5 - 15^2) = 0,202;$$

$$b = (57,5 \cdot 8,15 - 15 \cdot 26,975) / 62,5 = 1,024.$$

Напишемо шукане рівняння регресії: $Y = 0.202x + 1.024$.

Для того щоб отримати уявлення, наскільки добре обчислені за цим рівнянням значення Y_i узгоджуються з спостерігаються значеннями y_i знайдемо відхилення $Y_i - y_i$. Результати обчислень наведені в табл.

x_i	Y_i	y_i	$Y_i - y_i$
1,00	1,226	1,25	—0,024
1,50	1,327	1,40	—0,073
3,00	1,630	1,50	0,130
4,50	1,933	1,75	0,183
5,00	2,034	2,25	—0,216

Як видно з таблиці, не всі відхилення досить малі. Це пояснюється малим числом спостережень.

Кореляційна таблиця

При великій кількості спостережень одне і те ж значення x може зустрітися n_x раз, одне і те ж значення y - n_y раз, одна і та ж пара чисел (x, y) може спостерігатися n_{xy} раз. Тому дані спостережень групують, а саме. підраховують частоти n_x , n_y , n_{xy} . Все згруповані дані записують у вигляді таблиці, яку називають кореляційної.

Пояснимо пристрій кореляційної таблиці на прикладі

Y	x				
	10	20	30	40	n_y
0,4	5	-	7	14	26
0,6	-	2	6	4	12
0,8	3	19	-	-	22
n_x	8	21	13	18	$n=60$

У першому рядку таблиці вказані спостережувані значення (10; 20; 30; 40) ознаки X , а в першому стовчику - спостережувані значення (0,4; 0,6; 0,8) ознаки Y . На перетині рядків і стовпців знаходяться частоти n_{xy} спостережуваних пар значень ознак. Наприклад, частота 5 вказує, що пара чисел (10; 0,4) спостерігалася 5 разів. Всі частоти поміщені в прямокутнику, сторони якого проведені жирними відрізками. Риска означає, що відповідна пара чисел, наприклад (20; 0,4), не спостерігалася.

В останньому стовпці записані суми частот рядків. Наприклад, сума частот першого рядка «жирного» прямокутника дорівнює $n_y = 5 + 7 + 14 = 26$; це число вказує, що значення ознаки Y , рівна 0,4 (в поєднанні з різними значеннями ознаки X), спостерігалася 26 разів.

В останньому рядку записані суми частот стовпців. Наприклад, число 8 вказує, що значення ознаки X , рівне 10 (в поєднанні з різними значеннями ознаки Y), спостерігалася 8 разів.

У клітці, розташованій в нижньому правому куті таблиці, поміщена сума всіх частот (загальне число всіх спостережень). Очевидно, $\sum n_x = \sum n_y = n$. У нашому прикладі

$$\sum n_x = 8+21+13+18=60 \text{ и } \sum n_y = 26+12+22 = 60.$$

Видишування параметрів вибіркового рівняння прямої лінії регресії по згрупованим даним

Для визначення параметрів рівняння прямої лінії регресії Y на X була отримана система рівнянь

$$(\sum x^2)\rho + (\sum x)b = \sum xy; (\sum x)\rho + nb = \sum y.$$

Скористаємося тотожностями:

$$\sum x = n\bar{x} \text{ (следствие из } \bar{x} = \sum x / n \text{);}$$

$$\sum y = n\bar{y} \text{ (следствие из } \bar{y} = \sum y / n \text{);}$$

$$\sum x^2 = n\bar{x}^2 \text{ (следствие из } \bar{x}^2 = \sum x^2 / n \text{),}$$

$$\sum xy = \sum n_{xy}xy \text{ (учтено, что пара чисел } (x, y) \text{ наблюдалась } n_{xy} \text{ раз).}$$

Підставивши праві частини тотожностей в систему і скоротивши обидві частини другого рівняння на n , отримаємо

$$\begin{aligned} (n\bar{x}^2)\rho_{yx} + (n\bar{x})b &= \sum n_{xy}xy, \\ (\bar{x})\rho_{yx} + b &= \bar{y}. \end{aligned}$$

Вирішивши цю систему, знайдемо параметри ρ_{yx} і b .

Однак доцільніше, ввівши нову величину – вибіркового коефіцієнт кореляції, написати рівняння регресії в іншому вигляді.

$$\bar{y}_x - \bar{y} = \rho_{yx} (x - \bar{x}).$$

Знайдемо з системи коефіцієнт регресії, враховуючи, що $\bar{x}^2 - (\bar{x})^2 = \bar{\sigma}_x^2$

Помножимо обидві частини рівності на дріб $\bar{\sigma}_x / \bar{\sigma}_y$ отримаємо

$$\rho_{yx} \cdot \frac{\bar{\sigma}_x}{\bar{\sigma}_y} = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\bar{\sigma}_x\bar{\sigma}_y}.$$

Позначимо праву частину рівності через r_g і назовемо її вибіркоvim коефіцієнтом кореляції

$$r_g = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\bar{\sigma}_x\bar{\sigma}_y}, \quad \rho_{yx} = r_g \frac{\bar{\sigma}_y}{\bar{\sigma}_x}.$$

Остаточно отримаємо вибіркoве рівняння прямої лінії регресії Y на X виду

$$\bar{y}_x - \bar{y} = r_g \frac{\bar{\sigma}_y}{\bar{\sigma}_x} (x - \bar{x}).$$

Вибірковий коефіцієнт кореляції є оцінкою коефіцієнта кореляції

$$r = \frac{\mu_{xy}}{\sigma_x \sigma_y} = \frac{M(XY) - M(X) \cdot M(Y)}{\sigma_x \sigma_y}.$$

Вибірковий коефіцієнт кореляції

Як впливає з попереднього параграфа, вибіркoвий коефіцієнт кореляції визначається рівністю

$$r_g = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\bar{\sigma}_x\bar{\sigma}_y}.$$

Якщо величини Y і X незалежні, то коефіцієнт кореляції $r = 0$; якщо, $\rho = \pm 1$, то Y і X зв'язані лінійною функціональною залежністю. Звідси впливає, що коефіцієнт кореляції r вимірює силу (тісноту) лінійного зв'язку між Y і X .

Вибірковий коефіцієнт кореляції r_g є оцінкою коефіцієнта кореляції генеральної сукупності і тому також служить для вимірювання лінійного зв'язку між величинами – кількісними ознаками Y і X . Припустимо, що вибіркoвий коефіцієнт кореляції, знайдений за вибіркою, виявився відмінним від нуля. Так як вибірка відібрана випадково, то звідси ще не можна зробити висновок, що коефіцієнт кореляції генеральної сукупності також різниться від нуля. Виникає необхідність перевірити гіпотезу про значущість (суттєвості) вибіркового коефіцієнта кореляції (або, що те ж, про рівність нулю коефіцієнта кореляції генеральної сукупності). Якщо гіпотеза про рівність нулю генерального коефіцієнта кореляції буде відкинута, то вибіркoвий коефіцієнт кореляції значущий, а величини X і Y корельовані; якщо гіпотеза прийнята, то вибіркoвий коефіцієнт кореляції незначущий, а величини X і Y НЕ корельовані.

Перевірку гіпотези про значущість вибіркового коефіцієнта кореляції для випадку нормальної кореляції розглянемо далі

Якщо вибірка має досить великий обсяг і добре представляє генеральну сукупність (репрезентативна), то висновок про тісноту лінійної залежності між ознаками, отриманими за даними вибірки, до певної міри може бути поширене і на генеральну сукупність. Наприклад, для оцінки коефіцієнта кореляції r_g нормально розподіленої генеральної сукупності (при $n \geq 50$) можна скористатися формулою

$$r_g - 3 \frac{1 - r_g^2}{\sqrt{n}} \leq r_g \leq r_g + 3 \frac{1 - r_g^2}{\sqrt{n}}.$$

Методика обчислення вибіркового коефіцієнта кореляції

Нехай потрібно за даними кореляційної таблиці обчислити вибіркового коефіцієнта кореляції. Можна значно спростити розрахунок, якщо перейти до умовних варіантів (при цьому величина r_s не зміниться)

$$u_i = (x_i - C_1)/h_1 \text{ и } v_j = (y_j - C_2)/h_2.$$

В цьому випадку вибіркового коефіцієнта кореляції обчислюють за формулою

$$r_s = (\sum n_{uv}uv - n\bar{u}\bar{v}) / (n\bar{\sigma}_u\bar{\sigma}_v).$$

Приклад. За даними кореляційної таблиці знайти коефіцієнта кореляції

Y	X						n _y
	10	20	30	40	50	60	
15	5	7	—	—	—	—	12
25	—	20	23	—	—	—	43
35	—	—	30	47	2	—	79
45	—	—	10	11	20	6	47
55	—	—	—	9	7	3	19
n _x	5	27	63	67	29	9	n=200

Перейдемо до умовних варіантів, взявши $C_x=40$, $C_y=35$, $h_x=10$, $h_y=10$.

u	u						n _u
	-3	-2	-1	0	1	2	
-2	5	7	—	—	—	—	12
-1	—	20	23	—	—	—	43
0	—	—	30	47	2	—	79
1	—	—	10	11	20	6	47
2	—	—	—	9	7	3	19
n _u	5	27	63	67	29	9	n = 200

Тепер для обчислення шуканої суми $\sum n_{uv}uv$ складемо розрахункову табл

1. У кожній клітині, в якій частота $n_{uv} \neq 0$, записують у правому верхньому кутку добуток частоти n_{uv} на варіанту u . Наприклад, в правих верхніх кутах клітин першого рядка записані добутки: $5 \cdot (-3) = -15$; $7 \cdot (-2) = -14$.
2. Складають все числа, поміщені в правих верхніх кутах клітин одного рядка і їх суму записують в клітку цього ж рядка стовпця u . Наприклад, для першого рядка $U = -15 + (-14) = -29$.
3. Множимо варіанту v на U і отриманий добуток записують в останню клітку того ж рядка, а саме, в клітку стовпця uv . Наприклад, в першому рядку таблиці $v = -2$, $U = -29$; отже, $uv = (-2) \cdot (-29) = 58$.

4. Нарешті, склавши все числа стовпчика vU , отримують суму $\sum_v vU$, яка дорівнює шуканій сумі $\sum n_{uv}uv$. Наприклад, для табл. Маємо $\sum_v vU = 169$; отже, шукана сума $\sum n_{uv}uv = 169$.

u	u ч 1						U= = $\sum n_{uv}u$	uU
	-3	-2	-1	0	1	2		
-2	-15 5 -10	-14 7 -14	—	—	—	—	-29	58
-1	—	-40 20 -20	-23 23 -23	—	—	—	-63	63
0	—	—	-30 30 0	0 47 0	2 2 0	—	-28	0
1	—	—	-10 10 10	0 11 11	20 20 20	12 6 6	22	22
2	—	—	—	0 9 18	7 7 14	6 3 6	13	26
$V= =$ $\sum n_{uv}u$	-10	-34	-13	29	34	12		$\sum_v vU =$ =169
uV	30	68	13	0	34	"	$\sum_u uV$ ==169	Контроль

Оскільки при знаходженні r_e вже обчислені $\bar{u}$, $\bar{v}$, $\bar{\sigma}_u$, $\bar{\sigma}_v$, то доцільно користуватися формулами:

$$\bar{\sigma}_x = h_1 \bar{\sigma}_u, \bar{\sigma}_y = h_2 \bar{\sigma}_v, \bar{x} = \bar{u} h_1 + c_1, \bar{y} = \bar{v} h_2 + c_2.$$

Вибіркове кореляційне відношення

Для оцінки тісноти лінійної кореляційної зв'язку між ознаками у вибірці служить вибірковий коефіцієнт кореляції. Для оцінки тісноти нелінійної кореляційної зв'язку вводять нові зведені характеристики:

η_{yx} - вибіркове кореляційне відношення Y до X ;

η_{xy} - виборочное кореляційне відношення X до Y .

Вибірковим кореляційним відношенням Y до X називають відношення міжгрупового середнього квадратичного відхилення до загальної середньої квадратическому відхилення ознаки Y :

$$\eta_{yx} = \sigma_{\text{межгр}} / \sigma_{\text{общ}}, \text{ або } \eta_{yx} = \sigma_{\bar{y}_x} / \bar{\sigma}_y.$$

$$\text{де } \sigma_{\bar{y}_x} = \sqrt{D_{\text{мнжгр}}} = \sqrt{(\sum n_x (\bar{y}_x - \bar{y})^2) / n}; \quad \sigma_y = \sqrt{D_{\text{общ}}} = \sqrt{(\sum n_y (y - \bar{y})^2) / n},$$

де n -обсяг вибірки (сума всіх частот); n_x -частота значення x ознаки X ; n_y -частота значення y ознаки Y ; $\bar{y}$ -загальна середня ознаки Y ; $\bar{y}_x$ – умовна середня ознаки Y .

Аналогічно визначається вибіркове кореляційне відношення X до Y : $\eta_{xy} = \sigma_{\bar{x}_y} / \bar{\sigma}_x$.

Приклад. Знайти η_{yx} за даними кореляційної табл

y	X			
	10	20	30	n_y
15	4	28	6	38
25	6	—	6	12
n_x	10	28	12	$n=50$
$\bar{y}_x$	21	15	20	

Знайдемо загальну середню: $\bar{y} = (\sum n_y y) / n = (38 \cdot 15 + 12 \cdot 25) / 50 = 17,4$.

Знайдемо загальне середнє відхилення:

$$\bar{\sigma}_y = \sqrt{(\sum n_y (y - \bar{y})^2) / n} = \sqrt{[38(15 - 17,4)^2 + 12(25 - 17,4)^2] / 50} = 4,27.$$

Знайдемо міжгрупове середньоквадратичне відхилення:

$$\sigma_{\bar{y}_x} = \sqrt{(\sum n_x (\bar{y}_x - \bar{y})^2) / n} = \sqrt{[10(21 - 17,4)^2 + 28(15 - 17,4)^2 + 12(20 - 17,4)^2] / 50} = 2,73.$$

Шукане кореляційне відношення

$$\eta_{yx} = \sigma_{\bar{y}_x} / \bar{\sigma}_y = 2,73 / 4,27 = 0,64.$$

Властивості вибіркового кореляційного відношення

Оскільки η_{xy} має ті ж властивості, що й η_{yx} , перерахуємо властивості тільки вибіркового кореляційного відношення η_{yx} , яке далі для спрощення запису будемо позначати через η і для простоти мови називати «кореляційним відношенням».

Властивість 1. Кореляційне відношення задовольняє подвійній нерівності $0 \leq \eta \leq 1$.

Властивість 2. Якщо $\eta = 0$, то ознака Y з ознакою X кореляційної залежності не пов'язаний.

Властивість 3. Якщо $\eta = 1$, то ознака Y пов'язаний з ознакою X функціональною залежністю.

Властивість 4. Вибіркове кореляційне відношення менше абсолютної величини вибіркового коефіцієнта кореляції: $\eta \geq |r_s|$.

Властивість 5. Якщо вибіркове кореляційне відношення дорівнює абсолютній величині вибіркового коефіцієнта кореляції, то має місце точна лінійна кореляційна залежність.

Найпростіші випадки криволінійної кореляції

Наприклад, функції регресії Y на X можуть мати вигляд:

$y = Ax^2 + b x + c$ (параболічна кореляція другого порядку);

$y = Ax^3 + bx^2 + cx + d$ (параболічна кореляція третього порядку).

Для визначення виду функції регресії будують точки $(X;Y)$ і по їх розташуванню роблять висновок про приблизний вигляді функції регресії; при остаточному рішенні беруть до уваги особливості, що впливають із сутності розв'язуваної задачі.

Теорія криволінійної кореляції вирішує ті ж завдання, що і теорія лінійної кореляції (встановлення форми і тісноти кореляційної зв'язку). Невідомі параметри рівняння регресії шукають методом найменших квадратів. Для оцінки тісноти криволінійної кореляції служать вибіркові кореляційні відношення.

Обмежимося параболічною кореляцією другого порядку, припустивши, що дані n спостережень (вибірки) дозволяють вважати, що має місце саме така кореляція. В цьому випадку вибіркоче рівняння регресії Y на X має вигляд

$y = Ax^2 + b x + c$, де A, B, C -невідомі параметри.

Користуючись методом найменших квадратів, отримують систему лінійних рівнянь щодо невідомих параметрів

$$\begin{aligned}(\sum n_x x^4)A + (\sum n_x x^3)B + (\sum n_x x^2)C &= \sum n_x \bar{y}_x x^2 \\ (\sum n_x x^3)A + (\sum n_x x^2)B + (\sum n_x x)C &= \sum n_x \bar{y}_x x; \\ (\sum n_x x^2)A + (\sum n_x x)B + nC &= \sum n_x \bar{y}_x.\end{aligned}$$

Приклад. Знайти вибіркоче рівняння регресії Y на X виду $y = Ax^2 + b x + c$ за даними кореляційної табл.

Y	x			
	1	1,1	1.2	n_y
6	8	2	—	10
7	—	30	—	30
7,5	—	1	9	10
n_x	8	33	9	$n=50$
$\bar{y}_x$	6	6,73	7.5	

Складемо розрахункову таблицю

x	n_x	$\bar{y}_x$	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \bar{y}_x$	$n_x \bar{y}_x x$	$n_x \bar{y}_x x^2$
1	8	6	8	8	8	8	48	48	48
1,1	33	6,72	36,3	39,93	43,93	48,32	222,09	244,30	268,73
1,2	9	7,5	10,8	12,96	15,55	18,66	67,50	81	97,20
Σ	50	-	55,1	60,89	67,48	74,98	337,59	373,30	413,93

Підставивши числа (суми) нижнього рядка в систему, отримаємо

$$74,98A + 67,48B + 60,89C = 413,93;$$

$$67,48A + 60,89B + 55,10C = 373,30;$$

$$60,89A + 55,10B + n50 = 337,59.$$

Вирішивши цю систему, знайдемо: $A = 1,94$, $B = 2,98$, $C = 1,10$. Напишемо шукане рівняння регресії:

$$\bar{y}_x = 1,94x^2 + 2,98x + 1,10.$$

Легко переконатися, що умовні середні, обчислені за цим рівнянням, незначно відрізняються від умовних середніх кореляційної таблиці.

Поняття про множинної кореляції

Розглядалася кореляційний зв'язок між двома ознаками. Якщо ж досліджується зв'язок між кількома ознаками, то кореляцію називають множинною. У найпростішому випадку число ознак дорівнює трьом і зв'язок між ними лінійний:

$$z = ax + by + c.$$

У цьому випадку виникають завдання:

- 1) знайти за даними спостережень вибіркове рівняння зв'язку виду $z = ax + by + c$. Іншими словами. потрібно знайти коефіцієнти регресії A і B і параметр C ;
- 2) оцінити тісноту зв'язку між Z і обома ознаками X , Y ;
- 3) оцінити тісноту зв'язку між Z і X (при постійному Y), між Z і Y (при постійному X).

Перше завдання вирішується методом найменших квадратів, причому зручніше шукати рівняння зв'язку виду

$$z - \bar{z} = A(x - \bar{x}) + B(y - \bar{y}), \text{ де}$$

$$A = \frac{r_{xz} - r_{yz}r_{xy}}{1 - r_{xy}^2} \cdot \frac{\sigma_z}{\sigma_x}; \quad B = \frac{r_{yz} - r_{xz}r_{xy}}{1 - r_{xy}^2} \cdot \frac{\sigma_z}{\sigma_y}.$$

Тут r_{xz} , r_{yz} , r_{xy} коефіцієнти кореляції відповідно між ознаками X і Z , Y і Z , X і Y ;

σ_x , σ_y , σ_z – середні квадратичні відхилення.

Тіснота зв'язку ознаки Z з ознаками X , Y оцінюється вибірковим сукупним коефіцієнтом кореляції

$$R = \sqrt{\frac{r_{xz}^2 - 2r_{xy}r_{xz}r_{yz} + r_{yz}^2}{1 - r_{xy}^2}}, \text{ причому } 0 \leq R \leq 1.$$

Тіснота зв'язку між Z і X (при постійному Y), між Z і Y (при постійному X) оцінюється відповідно приватними вибірковими коефіцієнтами кореляції:

$$r_{xz(y)} = \frac{r_{xz} - r_{xy}r_{yz}}{\sqrt{(1 - r_{xy}^2)(1 - r_{yz}^2)}}; \quad r_{yz(x)} = \frac{r_{yz} - r_{xy}r_{xz}}{\sqrt{(1 - r_{xy}^2)(1 - r_{xz}^2)}}.$$

Ці коефіцієнти мають ті ж властивості і той же сенс, що і звичайний вибірковий коефіцієнт кореляції, а саме, служать для оцінки лінійного зв'язку між ознаками.

Коефіцієнти рангової кореляції Кендалла і Спірмена

Коефіцієнти рангової кореляції Кендалла і Спірмена застосовується для виявлення взаємозв'язку між кількісними або якісними показниками/ознаками, якщо їх можна ранжувати.

1. Для визначеності будемо розташовувати об'єкти в порядку погіршення якості. Нехай є дві вибірки (наприклад, порівнюємо 2 ознаки) $A = \{x_i\}$ і $B = \{y_i\}$ однакового об'єму. Спочатку впорядкуємо об'єкти вибірки А і кожному значенню припишемо ранг рівний його порядковому номеру, аналогічно впорядкуємо вибірку В. Далі запишемо ранги впорядкованої множини А, а під ними ранги відповідних значень множини В. Позначимо $d_i = x_i - y_i$ різницю відповідних рангів, тоді вибірковий коефіцієнт кореляції спірмена знаходиться за формулою:

$$\rho_b = 1 - \frac{6 \sum d_i^2}{n^3 - n}, \text{ де } n - \text{об'єм вибірок. } |\rho_b| \leq 1$$

2. Впорядкувавши ранги двох вибірок як і в попередньому випадку будемо підраховувати R_i – кількість рангів більших за y_i . Тоді вибірковий коефіцієнт рангової кореляції Кендалла обчислюється за формулою: $\tau_b = \frac{4R}{n(n-1)} - 1$, де n – об'єм вибірки, $R = R_1 + R_2 + \dots + R_{n-1}$

Приклад. При прийомі на роботу семи кандидатів на вакантні посади було запропоновано два тести. Результати тестування (в балах) наведено в таблиці:

Тест	Кандидат						
	1	2	3	4	5	6	7
1	31	82	25	26	53	30	29
2	21	55	8	27	32	42	26

Обчислити рангові коефіцієнт кореляції Кендалла між результатами тестування по двох тестів і на рівні $\alpha = 0,05$ оцінити його значимість.

Рішення: для знаходження коефіцієнта кореляції Кендалла, розташуємо дані першого тесту по зростанню і визначимо число інверсій для даних другого тесту: перший елемент другого ряду (8) менше всіх наступних елементів, тобто спостерігається порядок, другий елемент другого ряду (27) перевищує третій і п'ятий елементи, тобто спостерігаються дві інверсії. Оскільки, починаючи з другого елементу, маємо ранжування (27, 26, 42, 21, 32, 55). З пар чисел (перестановок) (27, 26), (27, 42), (27, 21), (27, 32) і (27, 55) інверсії (порушення порядку, коли більше число стоїть зліва від меншого) є у першій і третій парі, то число інверсій дорівнює 2 і даний результат відзначений в останньому рядку. Для третього елемента (26), оскільки він перевищує тільки п'ятий елемент, спостерігається одна інверсія, що і зазначено нижче в останньому рядку.

Кандидати	3	4	7	6	1	5	2
№ теста							
1	25	26	29	30	31	53	82
2	8	27	26	42	21	32	55
K=5	-	2	1	2	-	-	-

Четвертий елемент (42) перевищує п'ятий і шостий елементи другого ряду, тому число інверсій для нього дорівнює двом. Далі п'ятий, шостий і сьомий елементи розташовані в порядку зростання, тобто для них є порядок - інверсій немає. Таким чином, загальне число інверсій (K) для другого ряду дорівнює 5. Коефіцієнт кореляції Кендалла дорівнює

$$\tau = 1 - \frac{4K}{n(n-1)} = 1 - \frac{4 \cdot 5}{7 \cdot 6} = 0,524$$

Коефіцієнт кореляції Кендалла значущий на рівні $\alpha = 0,05$, якщо значення статистики

$$t = \tau \sqrt{\frac{9n(n-1)}{2(2n+5)}} \quad t_{\phi} = 1,65$$

По таблиці функції Лапласа $t_{0,95} = 1,96$. Фактичне значення менше критичного, отже коефіцієнт кореляції не є значущим.

Приклад. При дегустації двох сортів вина семи експертами виставлені наступні бальні оцінки

Сорт	Експерты						
	1	2	3	4	5	6	7
Сорт А	2.6	5.3	2.9	2.5	3.1	8.2	3.0
Сорт В	2.7	3.2	2.6	0.8	2.1	5.5	4.2

Обчислити рангові коефіцієнт кореляції Спірмена між результатами експертної оцінки за двома сортами і на рівні $\alpha = 0,05$ оцінити його значимість.

Рішення: обчислення коефіцієнт кореляції Спірмена проведемо в Mathcad.

Mathcad calculations for Spearman's rank correlation coefficient:

Input data matrix ZS :

$$ZS := \begin{pmatrix} 2.6 & 5.3 & 2.9 & 2.5 & 3.1 & 8.2 & 3.0 \\ 2.7 & 3.2 & 2.6 & 0.8 & 2.1 & 5.5 & 4.2 \end{pmatrix}$$

Number of experts $n = 7$, indices $i := 1..7$

Calculation of differences $Z_i := ZS_{1,i} - ZS_{2,i}$:

Z_i	$(Z_i)^2$
-0.1	0.01
2.1	4.41
0.3	0.09
1.7	2.89
1	1
2.7	7.29
-1.2	1.44

Sum of squared differences: $\sum_{i=1}^7 (Z_i)^2$

Final result: $\rho = 0.694$

При перевірці значимості ρ виходять з того, що в разі справедливості нульової гіпотези про відсутність кореляційної зв'язку між змінними статистика

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}}$$

має t - розподіл Стюдента з $k = n - 2$ ступенями свободи. Тому ρ значущий на рівні α , якщо фактичне значення t_{ϕ} буде більше критичного (по абсолютній величині) $|t| > t_{1-\alpha, k}$, де $t_{1-\alpha, k}$ - табличне значення t - критерію Стюдента з $k = n - 2$ ступенями свободи на рівні значущості $\alpha = 0,05$. Згідно співвідношенню, фактичне значення $t_{\phi} = 2,16$, а табличне значення t - критерію

Стьюдента $t_{0.95,5} = 2,57$. Таким чином нульова гіпотеза не відкидається і коефіцієнт кореляції Спірмена незначущий.