

ЛЕКЦІЯ 7

КОРЕЛЯЦІЙНИЙ АНАЛІЗ

1. Поняття про кореляцію

Одним із найважливіших завдань будь-якого дослідження, зокрема і психолого-педагогічного, є встановлення зв'язку між величинами або факторами, зміна яких визначає сутність процесу, що вивчається. Щоб пізнати досліджуване явище, треба вивчити не тільки його зв'язки з навколишніми явищами (факторами), але й також взаємозв'язки всіх його сторін, тобто треба встановити закономірності змін взаємопов'язаних явищ і показників, що їх характеризують.

В роботі практичного психолога часто необхідно аналізувати залежність між двома або декількома змінними величинами (ознаками). Якщо дві деякі характеристики, отримані для одного і того ж «об'єкта», мають тенденцію змінюватися сумісно так, що створюється можливість передбачити одну з них за значенням іншої, то кажуть, що ці характеристики **корелюють** одна з одною. Відповідно в статистиці кореляція виражає ступінь взаємозв'язку між такими характеристиками. Кількісно ця ступінь взаємозв'язку виражається за допомогою **коефіцієнта кореляції**.

Кореляційний зв'язок – це узгоджені зміни двох ознак або більшої кількості ознак (множинна кореляційний зв'язок).

Кореляційний зв'язок відображає той факт, що мінливість однієї ознаки знаходиться в деякій відповідності до мінливості іншої. З іншого боку, кореляційний зв'язок може говорити не про залежність ознак між собою, а про залежність цих ознак від іншої (інших). Кореляційні зв'язки не може розглядатися як свідчення причинно-наслідкового зв'язку, вони свідчать лише про те, що зміни однієї ознаки, як правило, супроводжують певними змінами іншої, але чи знаходиться причина змін в одній з ознак або вона виявляється за межами досліджуваної пари ознак, нам невідомо.

Кореляційна залежність – це зміни, які вносять значення однієї ознаки в ймовірність появи різних значень іншої ознаки.

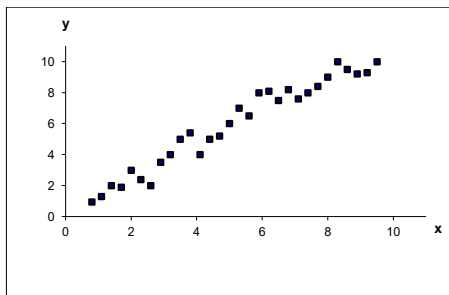
Кореляційні зв'язки розрізняються за *формою, напрямком і ступенем (силою)*.

За формою кореляційний зв'язок може бути лінійної і криволінійної.

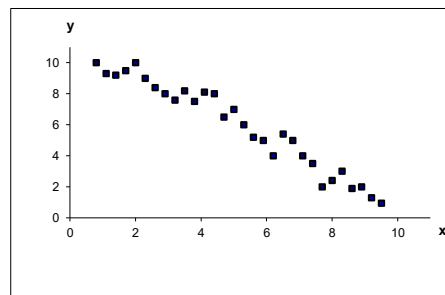
За формою кореляційний зв'язок може бути *лінійним і криволінійним*.

Лінійний зв'язок – якщо зі збільшенням або зменшенням однієї змінної, друга змінна в середньому або також зростає, або зменшується (рис. 7.1, 7.2).

Наприклад, прямолінійною можна назвати зв'язок між кількістю тренувань на тренажері і кількістю правильно вирішуваних завдань у контрольній сесії.

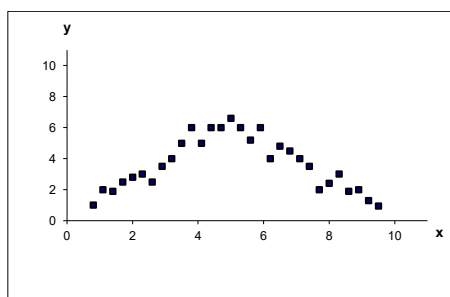


7.1. Лінійний позитивний зв'язок



7.2. Лінійний негативний зв'язок

Криволінійною може бути, наприклад, зв'язок між рівнем мотивації і ефективністю виконання завдання. При підвищенні мотивації ефективність виконання завдання спочатку зростає, потім досягається оптимальний рівень мотивації, якому відповідає максимальна ефективність виконання завдання; подальшого підвищення мотивації супроводжує вже зниження ефективності (рис.7.3).



7.3. Криволінійний зв'язок

За напрямком кореляційний зв'язок може бути позитивним («прямий») і негативним («обернений»).

При позитивній (прямій) кореляції більш високих значень однієї ознаки відповідають більш високим значенням іншої, а більш низьким значенням однієї ознаки - низькі значення іншої (рис. 7.1).

При негативній (оберненій) кореляції високі значення однієї ознаки відповідають більш низьким значенням іншої (рис. 7.2).

Ступінь кореляційного зв'язку визначається за величиною коефіцієнта кореляції, що позначають часто як r . Величина коефіцієнта кореляції знаходиться в діапазоні від -1 до +1.

Сила зв'язку не залежить від його спрямованості і визначається за абсолютним значенням коефіцієнта кореляції.

Якщо коефіцієнт кореляції за модулем виявляється близьким до 1, то це відповідає високому рівню зв'язку між змінними.

Класифікація сили кореляції

Використовуються кілька систем класифікації сили кореляції.

Загальна класифікація засвідчує, що кореляція:

✓ **сильна**, або **тісна** при $|r| \geq 0,7$;

- ✓ *середня* при $0,5 \leq |r| < 0,7$;
- ✓ *помірна* при $0,3 \leq |r| < 0,5$;
- ✓ *слабка* при $0,2 \leq |r| < 0,3$;
- ✓ *дуже слабка* при $|r| < 0,2$;

Ця класифікація орієнтована на величину коефіцієнта кореляції і жодною мірою не реагує на рівень його значущості.

Класифікація кореляційних зв'язків:

- ✓ *висока* значуща кореляція – при r , що відповідає рівню статистичної значущості $p \leq 0,01$;
- ✓ *значуща* кореляція – при r , що відповідає рівню статистичної значущості $p \leq 0,05$;
- ✓ *тенденція достовірного* зв'язку – при r , що відповідає рівню статистичної значущості $p \leq 0,10$;
- ✓ *незначна* кореляція при r , що не досягає рівня статистичної значущості.

Змінні можуть бути виміряні в різних шкалах, саме це визначає вибір відповідного коефіцієнта кореляції. У таблиці представлені співвідношення між шкалами.

Тип шкали		Міра зв'язку
Змінна А	Змінна В	
Інтервальна або відношень	Інтервальна або відношень	r_{xy} – коефіцієнт лінійної кореляції Пірсона
Порядкова	Порядкова або інтервальна	R_s – коефіцієнт рангової кореляції Спірмена
Порядкова	Порядкова	τ – коефіцієнт кореляції Кендалла
Дихотомічна	Дихотомічна	ϕ коефіцієнт кореляції Пірсона
Дихотомічна	Порядкова	R_{rb} – рангово-бісеріальний коефіцієнт кореляції
Дихотомічна	Інтервальна або відношень	R_{bic} – бісеріальний коефіцієнт кореляції
Інтервальна	Порядкова	Не розроблений

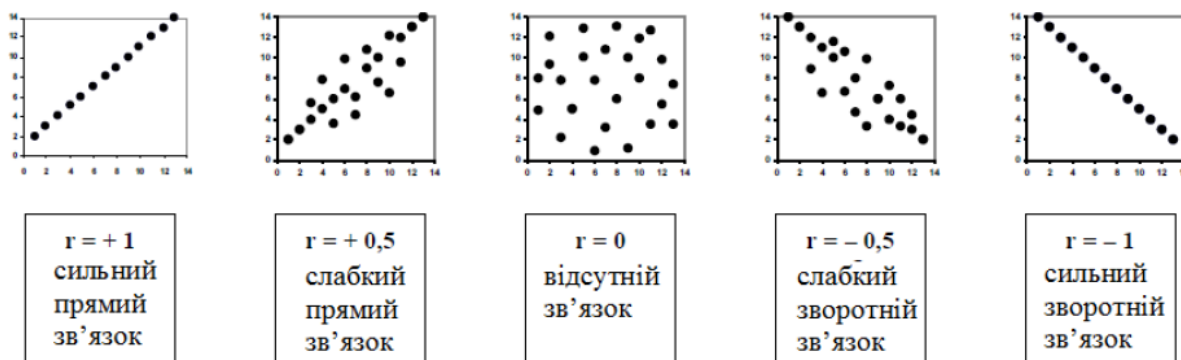
**порядкова шкала* вимірювань дозволяє ранжувати значення змінних. Вимірювання в порядковій шкалою містять інформацію тільки про порядок проходження величин, але не дозволяють сказати "наскільки одна величина більша за іншу", або "наскільки вона менше інший".

**рангова шкала.* У межах цієї шкали об'єкти розташовуються в порядку спадання чи зростання у них певної якості. При цьому кожній градації якості приписується свій порядковий номер (ранг). Фактично, об'єкти лише впорядковуються. Особливість шкали – однакові різниці між сусідніми рангами не означають однакової різниці між ступенями прояву виміряної якості.

При інтерпретації отриманого зв'язку велике значення має знак коефіцієнта кореляції. У випадку, коли коефіцієнт лінійної кореляції має знак «+», зв'язок між ознаками, що корелюються, має наступну особливість: більшій величині однієї ознаки

(змінної) відповідає більша величина іншої ознаки (іншої змінної). Іншими словами, якщо один показник (змінна) збільшується, то відповідно збільшується й інший показник (змінна).

Якщо ж коефіцієнт лінійної кореляції має знак «-», значить, більшій величині однієї ознаки відповідає меншій іншої. Тобто при наявності знаку «-» збільшення однієї змінної (властивості, значення) відповідає зменшенню іншої змінної. В даному випадку вибір змінної, у якій спостерігається тенденція до зростання, має довільний характер. Це може бути як змінна X, так і змінна Y. У випадку, коли експериментатор вважає, що збільшується змінна X, змінна Y відповідно буде зменшуватись, і навпаки. Ця інформація потрібна для правильної інтерпретації отриманої кореляційної залежності.



КОЕФІЦІЄНТ КОРЕЛЯЦІЇ ПІРСОНА

Умови для застосування коефіцієнта кореляції Пірсона

- 1) Змінні, що порівнюються повинні бути отримані в інтервальній шкалі або шкалі відносин.
- 2) Розподілу змінних X і Y повинні бути близькі до нормального.
- 3) Число ознак, що варіюють, у порівнянні зі іншими X і Y має бути однаковим.
- 4) Таблиці рівнів значущості для коефіцієнта кореляції Пірсона розраховані від $n=5$ до $n=1000$. Оцінка рівня значущості за таблицями здійснюється за числом ступенів свободи $v=n-2$.

Формула для підрахунку коефіцієнта кореляції така:

$$r_{xy} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{(n \sum x_i^2 - (\sum x_i)^2)(n \sum y_i^2 - (\sum y_i)^2)}}$$

де x_i – значення, що приймаються змінною X;

y_i – значення, що приймаються змінною Y.

ПРИКЛАД

20 школярам були дані тести на визначення домінування наочного або вербального мислення. Вимірювався середній час вирішення завдань тесту в секундах. Психолога

цікавить питання: чи існує взаємозв'язок між часом рішення наочно-образних завдань і вербальних завдань? Змінна X позначає середній час вирішення наочно-образних завдань, а змінна Y - середній час вирішення вербальних завдань тестів.

Розв'язання

Наведемо вихідні дані у вигляді таблиці, в якій введені додаткові стовпці, необхідні для розрахунку коефіцієнта кореляції Пірсона.

№ досліджуваних	Середній час розв'язання наочно-образних завдань	Середній час розв'язання вербальних завдань	$X \cdot Y$	$X \cdot X$	$Y \cdot Y$
	X	Y			
1	19	17	323	361	289
2	32	7	224	1024	49
3	33	17	561	1089	289
4	44	28	1232	1936	784
5	28	27	756	784	729
6	35	31	1085	1225	961
7	39	20	780	1521	400
8	39	17	663	1521	289
9	44	35	1540	1936	1225
10	44	43	1892	1936	1849
11	24	10	240	576	100
12	37	28	1036	1369	784
13	29	13	377	841	169
14	40	43	1720	1600	1849
15	42	45	1890	1764	2025
16	32	24	768	1024	576
17	48	45	2160	2304	2025
18	42	26	1092	1764	676
19	33	16	528	1089	256
20	47	26	1222	2209	676
Сума	731	518	20089	27873	16000

Гіпотези до задачі

H_0 : Кореляція між часом рішення наочно-образних завдань і вербальних завдань не відрізняється від нуля.

H_1 : Кореляція між часом рішення наочно-образних завдань і вербальних завдань статистично значимо відрізняється від нуля.

Розраховуємо емпіричну величину коефіцієнта кореляції Пірсона за формулою:

$$r_{xy} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{(n \sum x_i^2 - (\sum x_i)^2)(n \sum y_i^2 - (\sum y_i)^2)}}$$

$$r_{xy} = \frac{20 \cdot 20089 - 731 \cdot 518}{\sqrt{(20 \cdot 27873 - 731^2)(20 \cdot 16000 - 518^2)}}$$

$$= \frac{401780 - 378658}{\sqrt{(557460 - 534361)(320000 - 268324)}} = \frac{23122}{\sqrt{23099 \cdot 51676}} = \frac{23122}{34549} = 0,669$$

При знаходженні критичних значень для обчисленого коефіцієнта лінійної кореляції Пірсона число ступенів свободи розраховується як $\nu = n - 2$. В нашому випадку $n = 20$, тому $n - 2 = 20 - 2 = 18$. По таблиці критичних значень знаходимо:

$$r_{xy \text{ кр}} = \begin{cases} 0,44 & (p \leq 0,05) \\ 0,56 & (p \leq 0,01) \end{cases}$$

$k = n - 2$	P		$k = n - 2$	P	
	0,05	0,01		0,05	0,01
5	0,75	0,87	27	0,37	0,47
6	0,71	0,83	28	0,36	0,46
7	0,67	0,80	29	0,36	0,46
8	0,63	0,77	30	0,35	0,45
9	0,60	0,74	35	0,33	0,42
10	0,58	0,71	40	0,30	0,39
11	0,55	0,68	45	0,29	0,37
12	0,53	0,66	50	0,27	0,35
13	0,51	0,64	60	0,25	0,33
14	0,50	0,62	70	0,23	0,30
15	0,48	0,61	80	0,22	0,28
16	0,47	0,59	90	0,21	0,27
17	0,46	0,58	100	0,20	0,25
18	0,44	0,56	125	0,17	0,23
19	0,43	0,55	150	0,16	0,21
20	0,42	0,54	200	0,14	0,18
21	0,41	0,53	300	0,11	0,15
22	0,40	0,52	400	0,10	0,13
23	0,40	0,51	500	0,09	0,12
24	0,39	0,50	700	0,07	0,10
25	0,38	0,49	900	0,06	0,09
26	0,37	0,48	1000	0,06	0,09

«Вісь значимості»



$$r_{xy \text{ елт}} = 0,669$$

Висновок: з огляду на те, що величина розрахункового коефіцієнта кореляції потрапила в «зону значущості», гіпотеза H_0 відхиляється і приймається гіпотеза H_1 (при $p \leq 0,01$). Кореляція між часом рішення наочно-образних завдань і вербальних завдань статистично значимо відрізняється від нуля. Отримана позитивна залежність говорить про те, що чим вище середній час вирішення наочно-образних завдань, тим вище середній час вирішення вербальних завдань, і навпаки.

КОЕФІЦІЄНТ РАНГОВОЇ КОРЕЛЯЦІЇ СПІРМЕНА

Використовують для визначення щільності зв'язків між ознаками, якщо їх значення упорядковані або проранжовані за ступенем спадання або зростання ознаки. Коефіцієнт кореляції рангів розраховують за формулою:

$$R = 1 - \frac{6 \cdot \sum_{i=1}^n d^2}{n \cdot (n^2 - 1)}$$

$$R = 1 - \frac{6 \cdot \sum_{i=1}^n d^2}{n \cdot (n^2 - 1)}$$

n - обсяг вибірки;

d - різниця між рангами показників одних і тих самих реципієнтів

Умови для застосування коефіцієнта кореляції Спірмена

1) Змінні, що порівнюються повинні бути отримані в порядковій, інтервальній шкалах або в шкалі відносин.

2) Характер розподілу величин, корелюються не має значення.

3) Число ознак, що варіюють, повинно бути однаковим і перебувати в межах від 5 до 40.

4) У разі якщо є багато повторюваних рангів, то необхідно вносити поправку на однакові ранги.

$$T_a = \frac{(a_1^3 - a_1) + (a_2^3 - a_2) + \dots}{12},$$

$$T_b = \frac{(b_1^3 - b_1) + (b_2^3 - b_2) + \dots}{12},$$

де a_1, a_2, \dots – обсяг кожної групи однакових рангів у рангової ряду A ,

b_1, b_2, \dots – обсяг кожної групи однакових рангів у рангової ряду B .

За наявності однакових рангів:

$$R = 1 - 6 \cdot \frac{\sum d^2 + T_a + T_b}{n(n^2 - 1)}$$

Наприклад:

Значення	Ранг	Значення	Ранг
9	1,5	10	1
9	1,5	13	2
10	4	23	3,5
10	4	14	3,5
10	4	14	5
13	6,5	6	6
13	6,5	20	7
14	9	9	8,5
14	9	11	8,5
14	9	12	11
15	11	13	11
16	12,5	24	11
16	12,5	7	13
17	14	9	14
18	15,5	19	15
18	15,5	14	16
20	17	18	17

$$T_a = \frac{(2^3 - 2) + (3^3 - 3) + (2^3 - 2) + (3^3 - 3) + (2^3 - 2) + (2^3 - 2)}{12} = \frac{6 + 24 + 6 + 24 + 6 + 6}{12} = \frac{72}{12} = 6$$

$$T_b = \frac{(2^3 - 2) + (2^3 - 2) + (3^3 - 3)}{12} = \frac{6 + 24 + 6}{12} = \frac{36}{12} = 3$$

$$R = 1 - 6 \cdot \frac{767,75 + 6 + 3}{4896} = 1 - 0,95 = 0,05$$

Алгоритм розрахунку коефіцієнта рангової кореляції Спірмена

1. Визначити, які дві ознаки або дві ієрархії ознак братимуть участь в зіставленні як змінні А і Б.
2. Якщо це дві ознаки, проранжувати значення змінної А, присуджуючи ранг 1 найменшим значенням. Занести ранги в відповідний стовпець таблиці по порядку номерів досліджуваних.
3. Проранжувати значення змінної Б, відповідно до тих самих правил. Занести ранги в відповідний стовпець таблиці за порядком номерів досліджуваних або ознак

4. При наявності однакових показників присвоїти їм ранг, що представляє середню арифметичну співпадаючих рангів
5. Підрахувати різниці d між рангами А і Б по кожному рядку таблиці і занести їх до відповідного стовпця.
6. Піднести кожен різницю рангів до квадрату d^2 і внести отримані значення до відповідного стовпця.
7. Підрахувати суму квадратів рангів Σd^2 .
8. Розрахувати коефіцієнт рангової кореляції R за формулою
9. Визначити за таблицею Додатку 6 критичні значення $R_{\text{крит}}$ для даної кількості пар. Якщо $R_{\text{крит}}$ перевищує критичне значення або, принаймні дорівнює йому, кореляція статистично значуща.

Задача. У групи студентів були виміряні показники соціального інтелекту (композиторна оцінка), за тестом Гілфорда і показники рівня агресивності Ассингера. Чи існує зв'язок між цими особистісними якостями?

H_0 : Кореляція між показниками соціального інтелекту і рівня агресивності суттєво не відрізняється від 0 (є випадковою).

H_1 : Кореляція між показниками соціального інтелекту і рівня агресивності суттєво відрізняється від 0 (є не випадковою).

Соціальний інтелект		Рівень агресії		Різниця	Квадрат різниці
метричні значення	ранг	метричні значення	ранг	рангів (d)	рангів (d ²)
55	1	25	10	-9	81
52	2	16	14	-12	144
49	3	18	13	-10	100
46	4	20	11	-7	49
42	5	34	6	-1	1
39	6	35	5	1	1
38	7	38	3	4	16
37	8	40	2	6	36
35	9	36	4	5	25
34	10	14	15	-5	25
33	11	29	8	3	9
24	12	19	12	0	0
22	13	27	9	4	16
21	14	45	1	13	169
12	15	33	7	8	64
					$\Sigma d^2 = 736$

$$R = 1 - \frac{6 \cdot \sum_{i=1}^n d^2}{n \cdot (n^2 - 1)}$$

$$R = 1 - \frac{6 \cdot 736}{15(15^2 - 1)} = 1 - \frac{4416}{15 \cdot 224} = 1 - \frac{4416}{3360} = 1 - 1.314 \approx -0,0314$$

За таблицею критичних значень визначимо теоретичні значення:

$$R_{кр} = \begin{cases} 0,52 & \text{при } p = 0,05 \\ 0,66 & \text{при } p = 0,01 \end{cases}$$

n	P		n	P		n	P	
	0,05	0,01		0,05	0,01		0,05	0,01
5	0,94	—	17	0,48	0,62	29	0,37	0,48
6	0,85	—	18	0,47	0,60	30	0,36	0,47
7	0,78	0,94	19	0,46	0,58	31	0,36	0,46
8	0,72	0,88	20	0,45	0,57	32	0,36	0,45
9	0,68	0,83	21	0,44	0,56	33	0,34	0,45
10	0,64	0,79	22	0,43	0,54	34	0,34	0,44
11	0,61	0,76	23	0,42	0,53	35	0,33	0,43
12	0,58	0,73	24	0,41	0,52	36	0,33	0,43
13	0,56	0,70	25	0,49	0,51	37	0,33	0,43
14	0,54	0,68	26	0,39	0,50	38	0,32	0,41
15	0,52	0,66	27	0,38	0,49	39	0,32	0,41
16	0,50	0,64	28	0,38	0,48	40	0,31	0,40

Підкреслимо, що в таблиці критичних значень всі величини коефіцієнтів кореляції дані по абсолютній величині. Знак коефіцієнта враховується тільки при його інтерпретації.

Висновок: спостерігається помірна від'ємна кореляція між показниками соціального інтелекту та рівнем агресії.

ПРИКЛАД

Психолог з'ясує, як пов'язані між собою індивідуальні показники готовності до школи, отримані до початку навчання в школі, у 11 першокласників за 100-бальною шкалою і їх середня успішність в кінці навчального року за 5-бальною шкалою.

Розв'язання

Для вирішення завдання були проранжовано показники готовності до школи і середня успішність першокласників. Дані і результати вирішення занесені в таблицю.

Гіпотези до задачі

H_0 : Кореляція між показником готовності до школи і середньою успішністю в кінці навчального року не відрізняється від нуля.

H_1 : Кореляція між показником готовності до школи і середньою успішністю в кінці навчального року статистично значимо відрізняється від нуля.

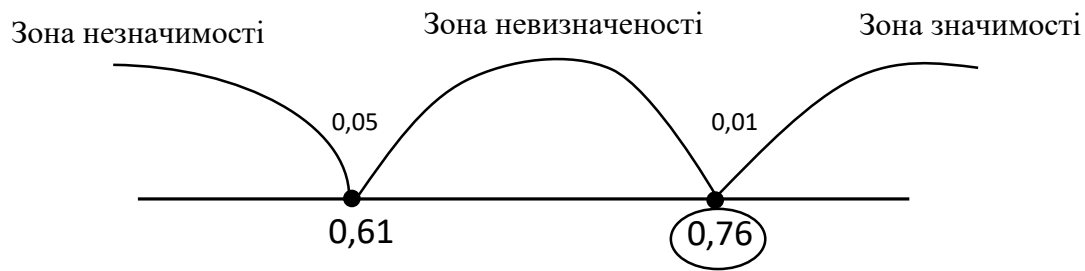
Досліджуваній	Показник готовності до школи		Середня успішність в кінці навчального року		d	d ²
	Значення	Ранг	Значення	Ранг		
1	65	3	4,3	2	1	1
2	70	5	4,7	7	-2	4
3	71	6	4,75	8	-2	4
4	50	1	4,35	3	-2	4
5	68	4	4,5	4	0	0
6	83	11	4,68	6	5	25
7	78	9	5	11	-2	4
8	62	2	3,5	1	1	1
9	75	8	4,9	10	-2	4
10	72	7	4,6	5	2	4
11	80	10	4,8	9	1	1
						$\sum = 52$

Значення емпіричного критерію знаходимо за формуло

$$r_{s \text{ эм}} = 1 - 6 \cdot \frac{\sum d^2}{n(n^2 - 1)} = 1 - 6 \cdot \frac{52}{11(11^2 - 1)} = 0,76.$$

Визначимо критичні значення r_s при $n = 11$ за таблицею:

$$r_{s \text{ кр}} = \begin{cases} 0,61 & (p \leq 0,05) \\ 0,76 & (p \leq 0,01) \end{cases}$$



Відповідь: $r_{s\text{ емп}} = 0,76$, нульова гіпотеза відхиляється і приймається альтернативна гіпотеза ($p=0,01$). Кореляція між показником готовності до школи і середньою успішністю в кінці навчального року відрізняється від нуля. Можна стверджувати, що показники шкільної готовності і підсумкові оцінки першокласників пов'язані позитивною кореляційною залежністю – чим вище показник шкільної готовності, тим краще вчиться першокласник.