КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Тема 5

- **Корреляция – вероятностная связь, взаимозависимость случайных величин**

- **Коэффициент корреляции —** это показатель степени связи между двумя переменными или измерениями. Обычно он обозначается буквой ***r***

- Коэффициент корреляции изменяется от -1 до +1

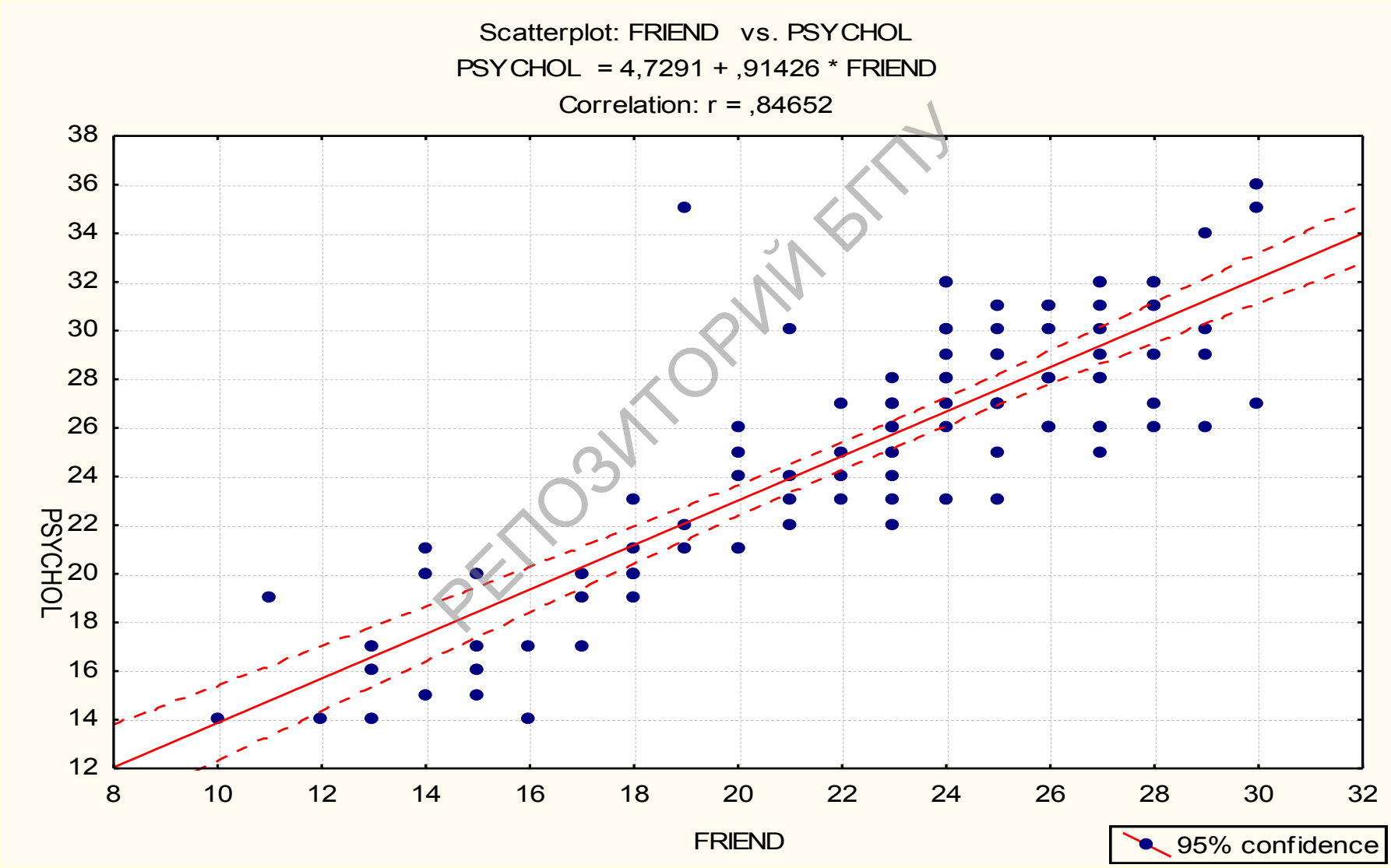
Величина коэффициента корреляции по модулю показывает степень зависимости

- r=0 — нет никакой связи;
- r =0.01–0.3 — слабая связь;
- r =0.31–0.7 — умеренная связь;
- r =0.71–0.99 — сильная связь;
- r =1 — совершенная связь.

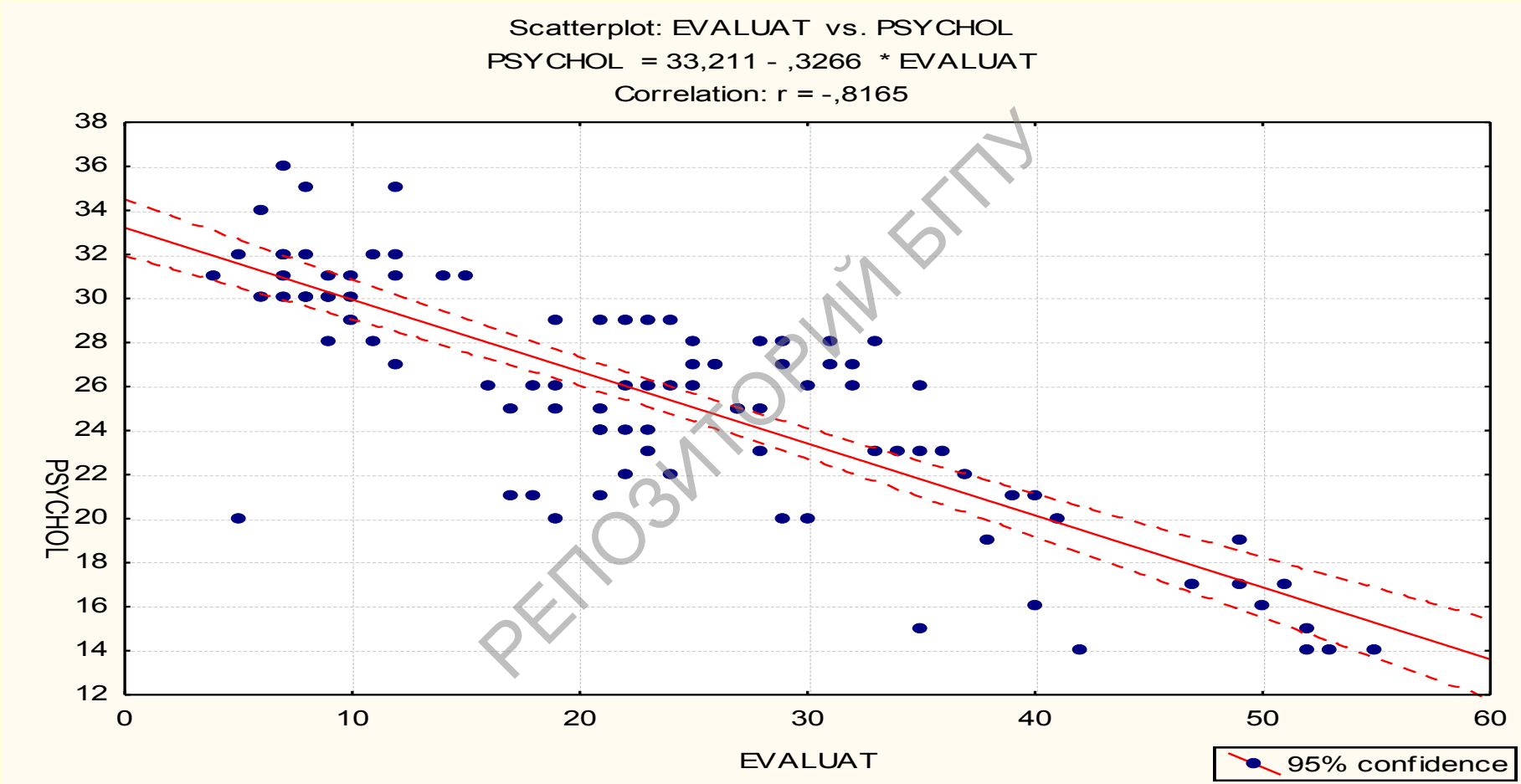
Знак коэффициента корреляции показывает направление отношений

- "+" — положительная (прямая) зависимость,
- "-" — отрицательная (обратная) зависимость

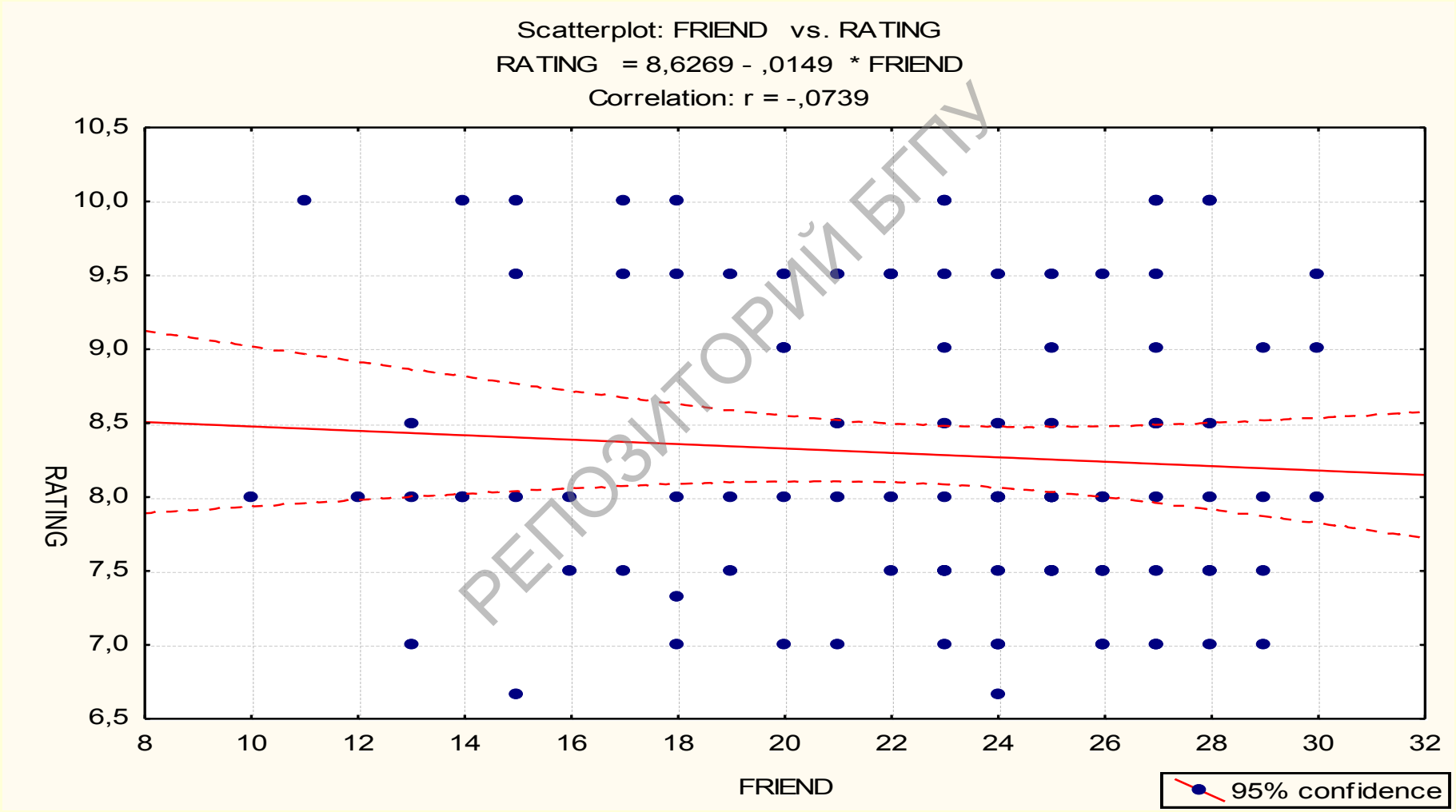
ПОЛОЖИТЕЛЬНАЯ КОРРЕЛЯЦИЯ



ОТРИЦАТЕЛЬНАЯ КОРРЕЛЯЦИЯ



ОТСУТСТВИЕ КОРРЕЛЯЦИИ



Значимость коэффициента корреляции

- H_0: вероятностная связь между рассматриваемыми свойствами генеральной совокупности отсутствует ($p>0,05$)

- H_1: имеется вероятностная связь между рассматриваемыми свойствами генеральной совокупности ($p\leq0,05$)

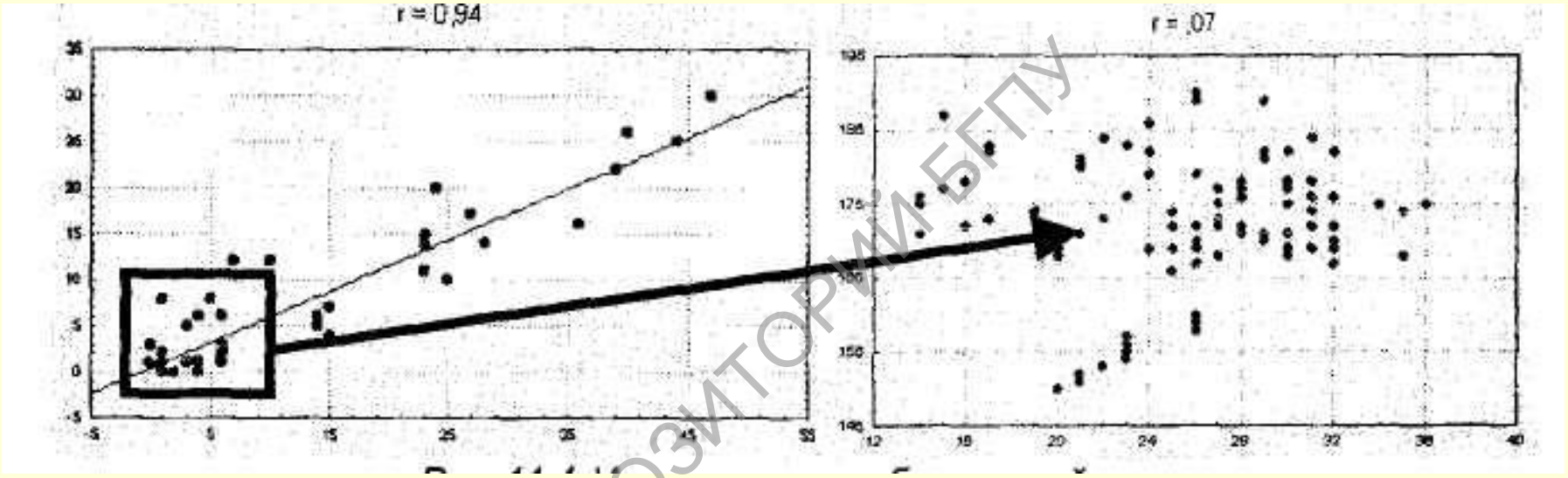
- При интерпретации коэффициента корреляции следует помнить, что существование даже очень большого коэффициента корреляции не гарантирует причинной связи между переменными

- Существование низкого коэффициента корреляции между некоторыми переменными не гарантирует отсутствия причинных связей между этими переменными

низкая корреляции при наличии причинной связи:

- Связь между явлениями не является линейной
- Действие неучтенных факторов, которые мешают проявиться зависимости
- Ограниченный интервал данных

Ограниченный интервал данных



Коэффициент ранговой корреляции Спирмена

- шкала порядка;
- объем обеих выборок $n \geq 5$;
- точность вычисления коэффициента корреляции Спирмена снижается при большом числе связанных рангов

Коэффициент ранговой корреляции Спирмена

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)},$$

D_i = *ранг* X_i — *ранг* Y_i (разность рангов),
n — количество ранжированных пар

Есть ли связь между температурой воздуха и количеством купленных пачек мороженого?

День недели	Температура воздуха, °С, X_i	Количество купленных пачек мороженого, Y_i	Ранг X_i	Ранг Y_i	Разность рангов D_i	Квадрат разности рангов D_i^2
Пн	7	1	2	1	1	1
Вт	4	3	1	2	-1	1
Ср	13	5	4	3	1	1
Чт	16	7	5	4	1	1
Пт	10	9	3	5	-2	4
Сб	22	11	7	6	1	1
Вс	19	13	6	7	-1	1
n=7						$\sum D_i^2=10$

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 10}{7(49 - 1)} = 1 - \frac{60}{336} = \frac{336 - 60}{336} = \frac{276}{336} = 0{,}821.$$

значимость коэффициента корреляции Спирмена

- В специальных таблицах находим критические значения коэффициента корреляции для уровней значимости p=0,05 и p=0, 01
- Если вычисленный коэффициент корреляции больше критического значения, то он является значимым при выбранном уровне статистической значимости **p**

Критические значения коэффициента корреляции рангов Спирмена

n	p	
	0,05	**0,01**
5	0,94	-
6	0,85	-
7	0,78	0,94
8	0,72	0,88
9	0,68	0,83
10	0,64	0,79
11	0,61	0,76
12	0,58	0,73
13	0,56	0,70
14	0,54	0,68
15	0,52	0,66
16	0,50	0,64

- r_s=0,821 > $r_{kp\ 0,05}$=0,78
- r_s=0,821 < $r_{kp\ 0,01}$=0,94

- **ВЫВОД:** наблюдается положительная, сильная, значимая (p=0,05) связь между температурой воздуха и количеством купленного мороженого

Коэффициент корреляции Пирсона

- шкала равных отношений или интервалов;
- объем выборок $n \geq 7$;
- оценка уровня значимости по таблицам осуществляется при числе степеней свободы $k=n-2$

Коэффициент корреляции Пирсона – формула средних отклонений

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{[\sum_{i=1}^{n}(X_i - \overline{X})^2][\sum_{i=1}^{n}(Y_i - \overline{Y})^2]}},$$

Коэффициент корреляции Пирсона – формула «сырых» данных

$$r_{xy} = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}}.$$

X_i — значения одной переменной;
Y_i — значения другой переменной;
n — число пар данных, взятых для анализа (число испытуемых)

Есть ли связь между температурой воздуха и количеством купленных пачек мороженого?

День недели	Температура воздуха, °С, X_i	Количество купленных пачек мороженого, Y_i	X_i^2	Y_i^2	$X_i \cdot Y_i$
Пн	7	1	49	1	7
Вт	4	3	16	9	12
Ср	13	5	169	25	65
Чт	16	7	256	49	112
Пт	10	9	100	81	90
Сб	22	11	484	121	242
Вс	19	13	361	169	247
n=7	$\sum X_i=91$	$\sum Y_i=49$	$\sum X_i^2=1435$	$\sum Y_i^2=455$	$\sum X_i \cdot Y_i=775$

$$r_{xy} = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2]}} =$$

$$= \frac{7 \cdot 775 - 91 \cdot 49}{\sqrt{[7 \cdot 1435 - 91^2][7 \cdot 455 - 49^2]}} = \frac{966}{\sqrt{1764 \cdot 784}} = 0{,}821.$$

значимость коэффициента корреляции Пирсона:

- Вычисляем количество степеней свободы k=n-2=7-2=5
- В специальных таблицах находим критические значения коэффициента корреляции для уровней значимости p=0,05 и p=0, 01
- Если вычисленный коэффициент корреляции больше критического значения, то он является значимым при выбранном уровне статистической значимости **p**

Критические значения коэффициента корреляции Пирсона

k=n-2	p	
	0,05	**0,01**
5	0,75	0,87
6	0,71	0,83
7	0,67	0,80
8	0,63	0,77

- $r_{xy}=0,821 > r_{kp\ 0,05}=0,75$
- $r_{xy}=0,821 < r_{kp\ 0,01}=0,87$

- **ВЫВОД:** наблюдается положительная, сильная, значимая (p=0,05) связь между температурой воздуха и количеством купленного мороженого