1. Корреляционный анализ

Изучение связей между переменными, интересует исследователя с точки зрения отражения соответствующих **причинно-следственных отношений**.

Корреляционная зависимость – это согласованные изменения двух (парная корреляционная связь) или большего количества признаков (множественная корреляционная связь). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой(-их) переменной(-ых).

Корреляционный анализ – статистический метод, позволяющий с использованием коэффициентов корреляции определить, существует ли зависимость между переменными и насколько она сильна.

Коэффициент корреляции – двумерная описательная статистика, количественная мера взаимосвязи (совместной изменчивости) двух переменных.

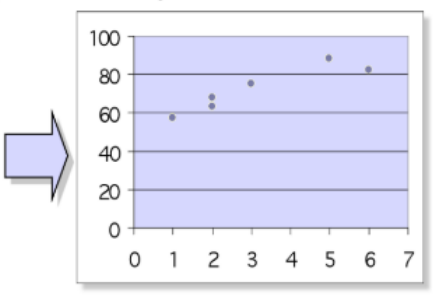
График рассеяния (Scatter Plot)

- Наглядное представление о связи двух переменных дает **график рассеяния**, на котором каждый объект представляет собой точку, координаты которой заданы значениями двух переменных. Таким образом, множество объектов представляет собой на графике множество точек. По конфигурации этого множества точек можно судить о характере связи между двумя переменными.
- Команда **«Графика»** → **«Рассеяния/Точки»**.

График рассеяния (Scatter Plot)

Пример: Рассматриваем две переменные: «Продолжительность подготовки (часов)» студентов перед экзаменом и «Итоговая оценка» (из 100 балов). Пытаемся визуально определить связь. Правда ли, что чем больше времени уделено подготовке, тем выше оценка? (Ответ на этот вопрос будет дан далее при расчете коэффициента корреляции Пирсона)

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75



Сила корреляции

- **Сила связи** не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции.
- **Коэффициент корреляции (r)** – это показатель, величина которого варьируется в пределах от –1 до +1.
- Если коэффициент корреляции равен 0, обе переменные линейно независимы друг от друга.

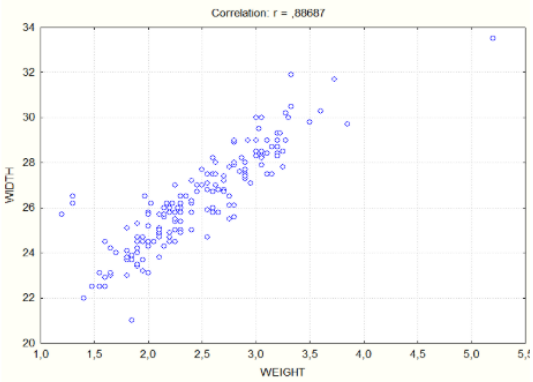
ЗНАЧЕНИЕ (по модулю)	ИНТЕРПРЕТАЦИЯ
до 0,2	очень слабая корреляция
до 0,5	слабая корреляция
до 0,7	средняя корреляция
до 0,9	высокая корреляция
свыше 0,9	очень высокая корреляция

Диаграмма рассеяния (Scatterplot, Scatter diagram)

Характеристики диаграммы:

- наклон (направление связи)
- ширина (сила, теснота связи)

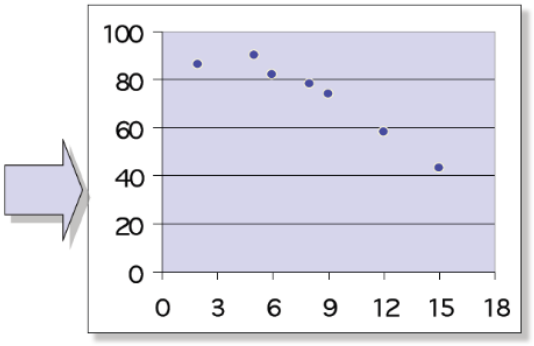
О силе связи можно судить по тому, насколько тесно расположены точки-объекты около линии регрессии - чем ближе точки к линии, тем сильнее связь.



Направление корреляции

Пример: На графике видно, что имеет место *отрицательная линейная зависимость*. Это означает, что увеличение переменной X приводит к уменьшению переменной Y.

Студент	Пропустил *x*	Оценка *y*
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78



2.1. Коэффициент корреляции Пирсона

Коэффициент корреляции *r*-Пирсона является мерой прямолинейной связи между переменными: его значения достигают максимума, когда точки на графике двумерного рассеяния лежат на одной прямой линии.

$$r = \frac{\sum z_{X_i} z_{Y_i}}{n - 1}$$

Пример: Исследование взаимосвязи веса и роста.

$$z_{X_i} = \frac{X_i - \bar{X}}{s_X}$$

стандартное отклонение для веса

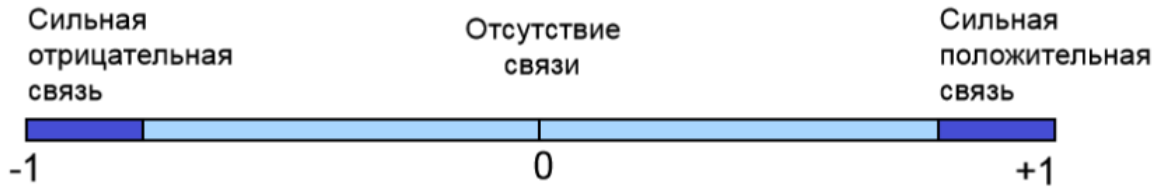
$$z_{Y_i} = \frac{Y_i - \bar{Y}}{s_Y}$$

стандартное отклонение для роста

для каждого X и Y (для каждого респондента)

	Вес	Рост
Дима	72	160
Гриша	66	144
Миша	68	154
Коля	74	210
Федя	68	182
Рома	64	159
	68,7	168,2

Интерпретация результатов



Значение r – Пирсона характеризует **уровень связи между переменными:**

- 0,75 – 1.00 очень высокая положительная
- 0,50 – 0.74 высокая положительная
- 0,25 – 0.49 средняя положительная
- 0,00 – 0.24 слабая положительная
- 0,00 – -0.24 слабая отрицательная
- -0,25 – -0.49 средняя отрицательная
- -0,50 – -0.74 высокая отрицательная
- -0,75 – -1.00 очень высокая отрицательная

Результаты коэффициента корреляции r – Пирсона для примера со студентами

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Студент	Часы x	Оценка y	xy	x^2	y^2
A	6	82	492	36	6724
B	2	63	126	4	3969
C	1	57	57	1	3249
D	5	88	440	25	7744
E	2	68	136	4	4624
F	3	75	225	9	5625
	Σx=19	**Σy=433**	**Σxy=1476**	**Σx^2=79**	**Σy^2=31935**

$$r = \frac{6 \cdot 1476 - 19 \cdot 433}{\sqrt{6 \cdot 79 - 19^2}\sqrt{6 \cdot 31935 - 433^2}} = 0{,}922$$

Оценка статистической значимости коэффициента корреляции

КРИТИЧЕСКОЕ ЗНАЧЕНИЕ *T*-КРИТЕРИЯ ОПРЕДЕЛЯЕТСЯ ИЗ ТАБЛИЦЫ ЗНАЧЕНИЙ *T*-РАСПРЕДЕЛЕНИЯ ДЛЯ ВЫБРАННОГО УРОВНЯ ЗНАЧИМОСТИ A И ЧИСЛА СТЕПЕНЕЙ СВОБОДЫ *DF=N–2*

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