

Дисциплина: «Основы научных исследований»

Лабораторная работа № 4

Исследование корреляционных зависимостей

1. Корреляция: определение, основные характеристики.
Корреляционный анализ.
2. Парные статистические связи.
3. Пример выполнения лабораторной работы №4.

1. Корреляционный анализ

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

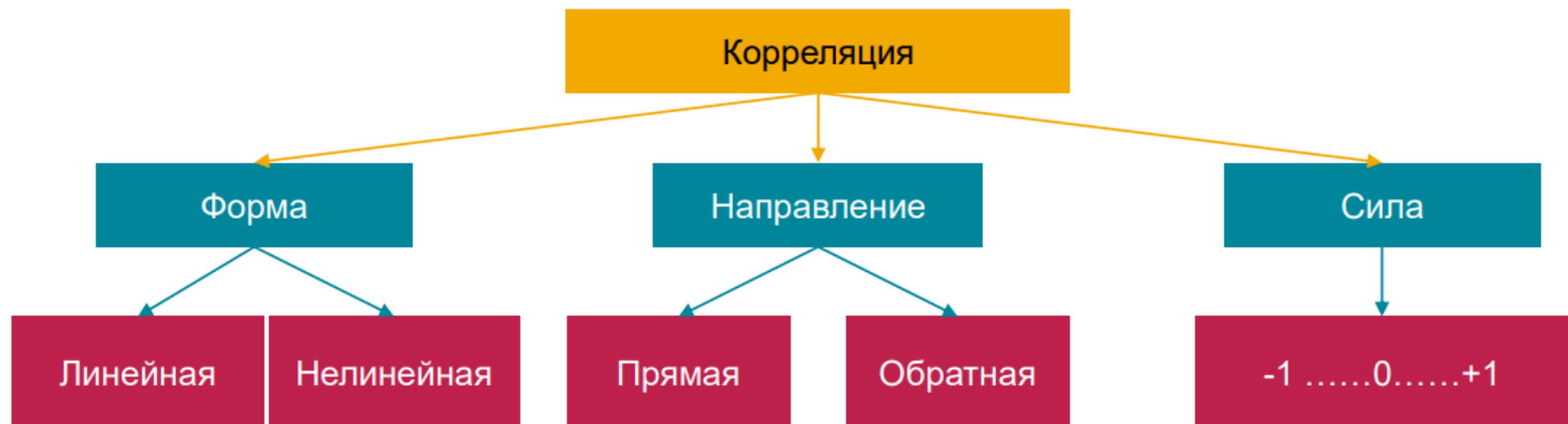
Корреляционная зависимость – это согласованные изменения двух (парная корреляционная связь) или большего количества признаков (множественная корреляционная связь). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой(-их)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

1. Корреляционный анализ

Характер связи между переменными



- При **положительной линейной корреляции** более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого.
- При **отрицательной линейной корреляции** более высоким значениям одного признака соответствуют более низкие значения другого, а более низким значениям одного признака – высокие значения другого.

1. Корреляционный анализ

Виды связи между переменными

1. **Прямая причинно-следственная связь** - переменная X определяет значение переменной Y.

Пример: Наличие воды ускоряет рост растений. Яд вызывает смерть. Температура воздуха прямо влияет на скорость таяния льда.

2. **Обратная причинно-следственная связь** - переменная Y определяет значение переменной X.

Пример: Исследователь может думать, что чрезмерное потребление кофе вызывает нервозность. Но, может быть, очень нервный человек выпивает кофе, чтобы успокоить свои нервы?

Виды связи между переменными

3. Связь, вызванная третьей (скрытой) переменной.

Пример: существует зависимость между числом утонувших людей и числом выпитых безалкогольных напитков в летнее время. Однако, обе переменные связаны с жарой и потребностью людей во влаге?

4. Связь, вызванная несколькими скрытыми переменными.

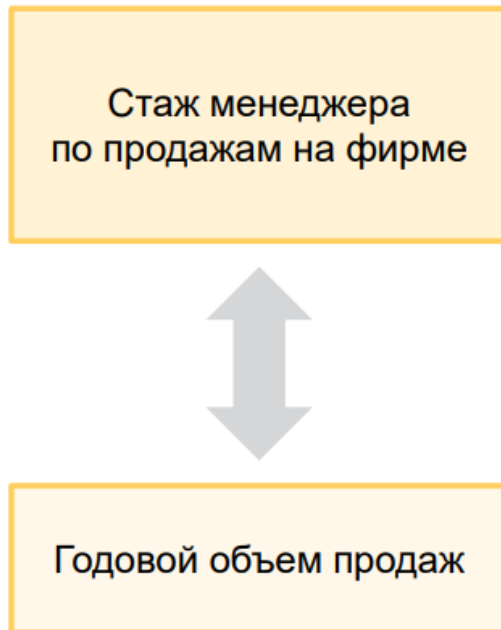
Пример: Исследователь может обнаружить значимую связь между оценками студентов в университете и оценками в школе. Но действуют и другие переменные: IQ, количество часов занятий, влияние родителей, мотивация, возраст, авторитет преподавателей.

5. Связи нет, наблюдаемая зависимость случайна.

Пример: Исследователь может найти связь между увеличением количества людей, которые занимаются спортом и увеличением количества людей, которые совершают преступления. Но здравый смысл говорит, что любая связь между этими двумя переменными является случайной.

Виды связи между переменными

Простая связь



Множественная связь

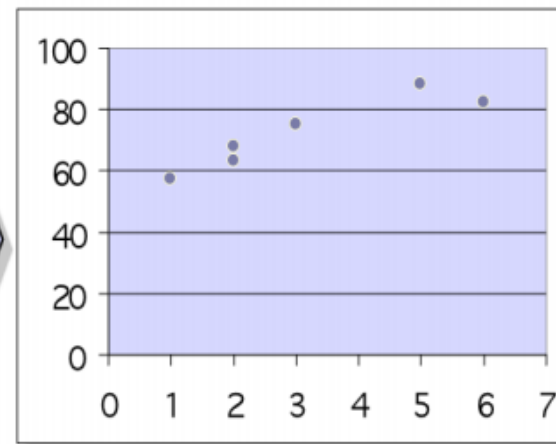
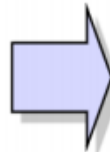


График рассеяния (Scatter Plot)

- 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.

Пример: Рассматриваем две переменные: «Продолжительность подготовки (часов)» студентов перед экзаменом и «Итоговая оценка» (из 100 баллов). 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

Сила корреляции

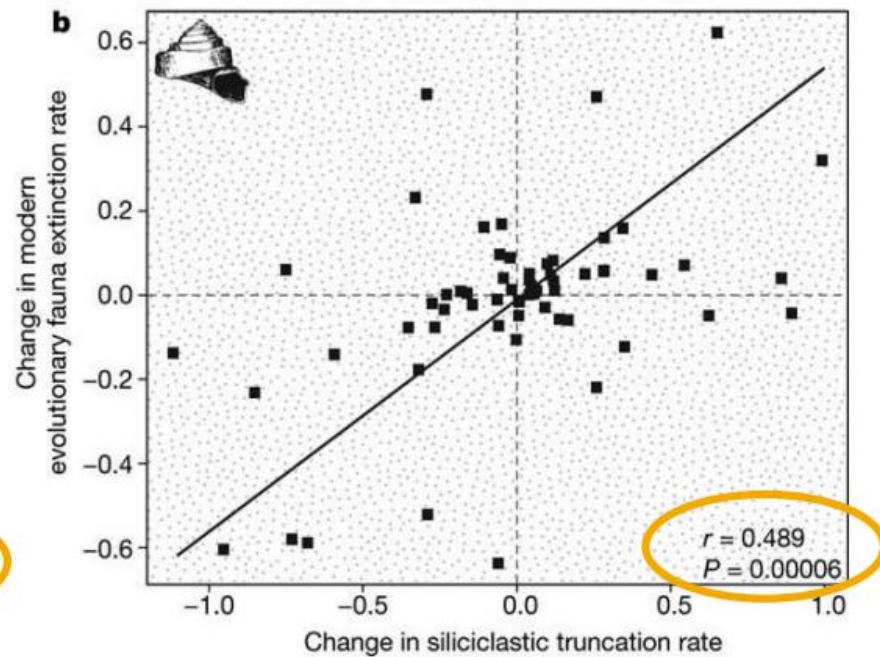
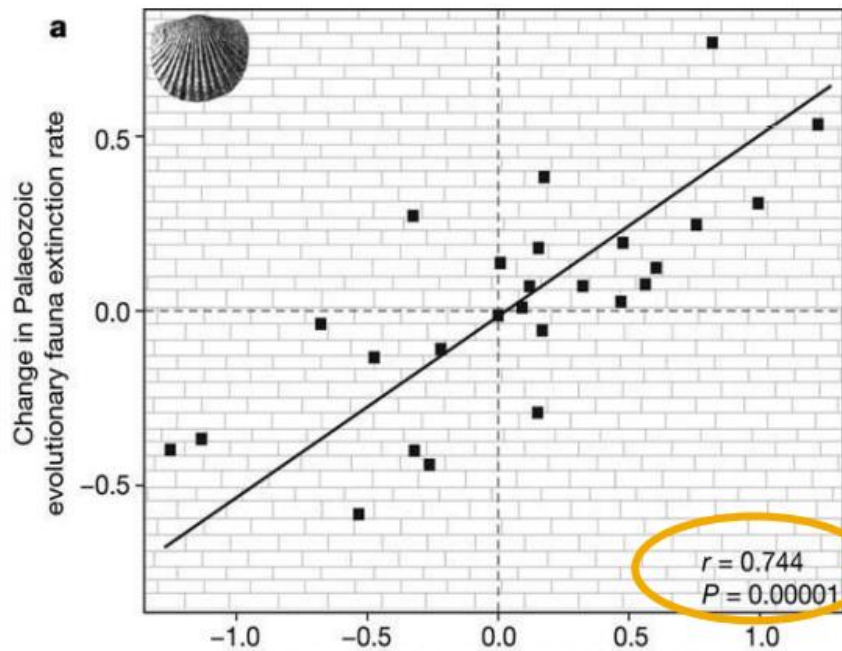
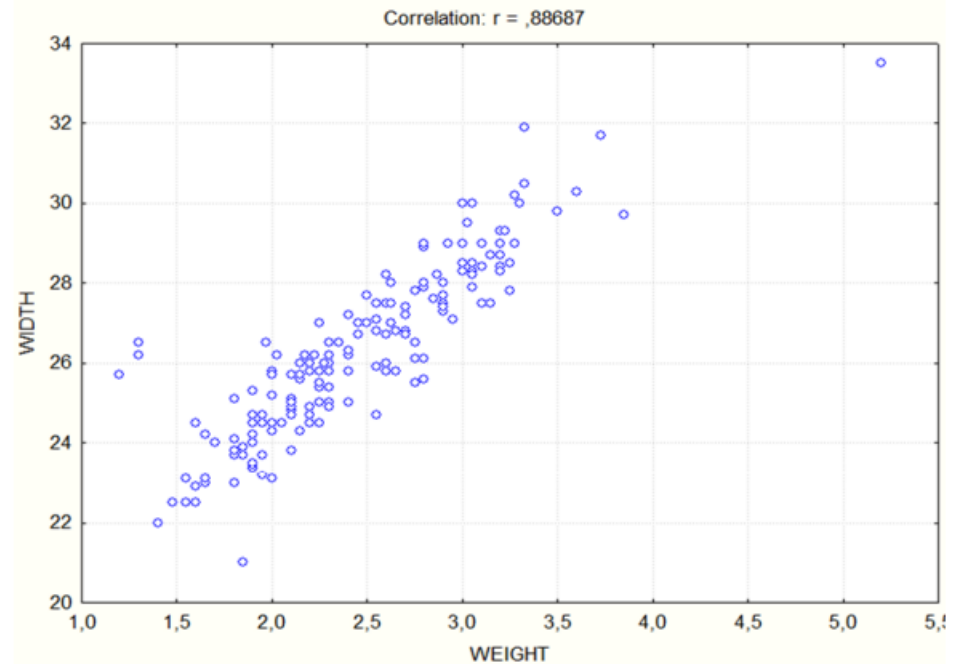


Диаграмма рассеяния (Scatterplot, Scatter diagram)

Характеристики диаграммы:

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

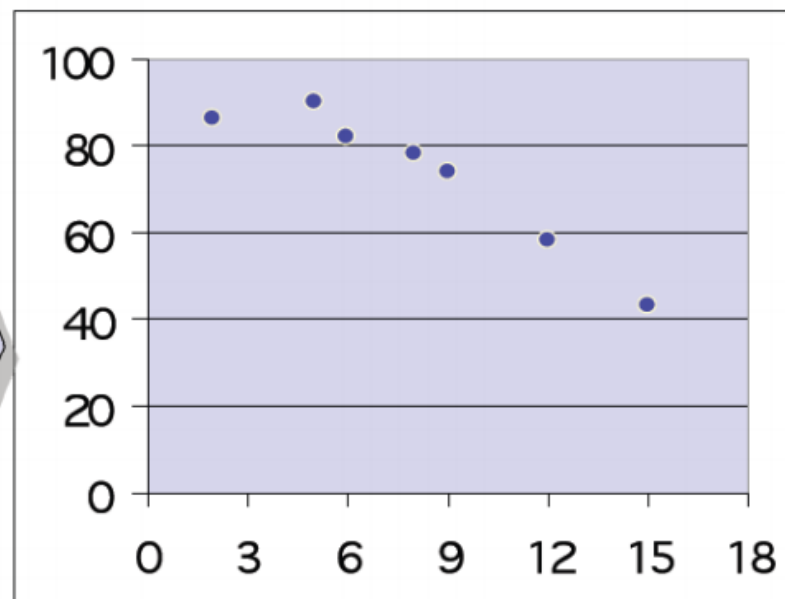
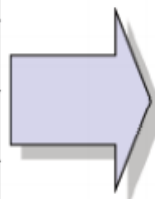
О силе связи можно судить по тому, насколько тесно расположены точки-объекты около линии регрессии - чем ближе точки к линии, тем сильнее связь.



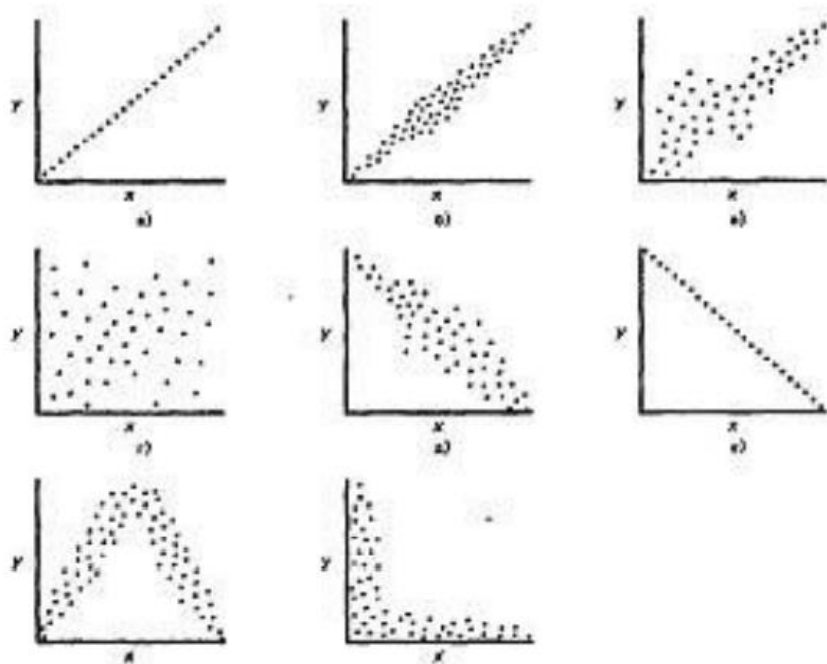
Направление корреляции

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



Примеры корреляций



- а) строгая положительная корреляция
- б) положительная корреляция
- в) слабая положительная корреляция
- г) нулевая корреляция

- д) отрицательная корреляция
- е) строгая отрицательная корреляция
- ж) нелинейная корреляция
- з) нелинейная корреляция

Ложная корреляция

- Если между двумя исследуемыми величинами установлена тесная зависимость, то из этого еще не следует их причинная взаимообусловленность. За счет эффектов одновременного влияния неучтенных факторов смысл истинной связи может искажаться. Поэтому такую корреляцию часто называют **«ложной»**.

Пример: «Аисты приносят детей»

Изучалась корреляция между числом аистов, свивших гнезда в южных районах Швеции, и рождаемостью в эти же годы в Швеции. Вычисления показали высокую положительную корреляцию между этими явлениями. Однако причинная зависимость не может быть выведена ни из какого наблюдаемого совместного изменения явлений. Оказалось, что одновременные синхронные изменения числа аистов и детей объясняются изменением среднего уровня жизни жителей Стокгольма. При исключении этой искажающей переменной прежней корреляции уже не наблюдалось.

- Для выявления «ложной» корреляции используются **частные корреляции**.

Частная корреляция

- Если две переменные коррелируют, всегда можно предположить, что эта корреляция обусловлена влиянием третьей переменной, как общей причины совместной изменчивости первых двух переменных.
- Для проверки этого предположения достаточно **исключить влияние этой третьей переменной** и вычислить корреляцию двух переменных без учета влияния третьей переменной (при фиксированных ее значениях).
- Корреляция, вычисленная таким образом называется **частной**.

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

2.1. Коэффициент корреляции Пирсона

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

2.1. Коэффициент корреляции Пирсона

Результаты коэффициента корреляции r – Пирсона для примера со студентами

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

Студент	Часы х	Оценка у	ху	х ²	у 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
	Σх=19	Σу=433	Σху=1476	Σх²=79	Σу²=31935

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

2.1. Коэффициент корреляции Пирсона

Оценка статистической значимости коэффициента корреляции

Критическое значение t -критерия определяется из таблицы значений t -распределения для выбранного уровня значимости α и числа степеней свободы $df=n-2$

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

2.1. Коэффициент корреляции Пирсона

ВАЖНО ЗАПОМНИТЬ!

- Коэффициент корреляции r - Пирсона оценивает только **линейную связь** переменных. Нелинейную связь данный коэффициент выявить не может.
- Коэффициент корреляции Пирсона очень чувствителен к **аутлаерам (выбросам)**.
- Корреляция **не подразумевает наличия причинно-следственной связи** между переменными.
- **Нельзя путать** коэффициент корреляции Пирсона с критерием Пирсона χ^2 -квадрат.

Задание на лабораторную работу №4:

По результатам восьми опытов определить коэффициент корреляции Пирсона для переменных: X – уровень внутреннего шума в кабине автомобиля (дБ) и Y – скорость движения автомобиля (м/с).

Сделать вывод об уровне связи между переменными.

№ опыта	X , дБ	Y , м/с
1	80+K	10+K
2	92+K	25+K
3	85+K	14+K
4	88+K	20+K
5	79+K	8+K
6	90+K	22+K
7	91+K	24+K
8	95+K	30+K

Примечание: K - номер студента по журналу.