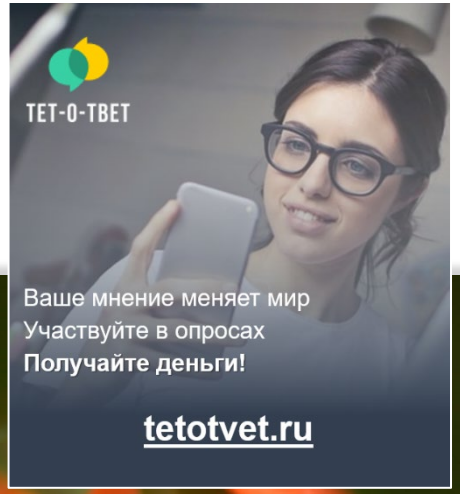
НАФИ



ТЕМА 6

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Для свободного использования в образовательных целях
Copyright © Академия НАФИ. Москва
Все права защищены
nafi.ru

ОГЛАВЛЕНИЕ

1. **Корреляция: определение, основные характеристики. Корреляционный анализ**

2. **Парные статистические связи**
   - 2.1 Коэффициент корреляции Пирсона (метрические шкалы)
   - 2.2 Коэффициент ранговой корреляции Спирмена (неметрические шкалы)
   - 2.3 Коэффициент ранговой корреляции Кендалла
   - 2.4 Коэффициент ранговой корреляции Гудмена - Краскела

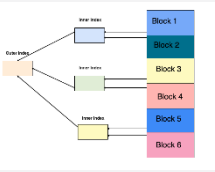
ЗАЧЕМ НУЖЕН КОРРЕЛЯЦИОННЫЙ АНАЛИЗ?

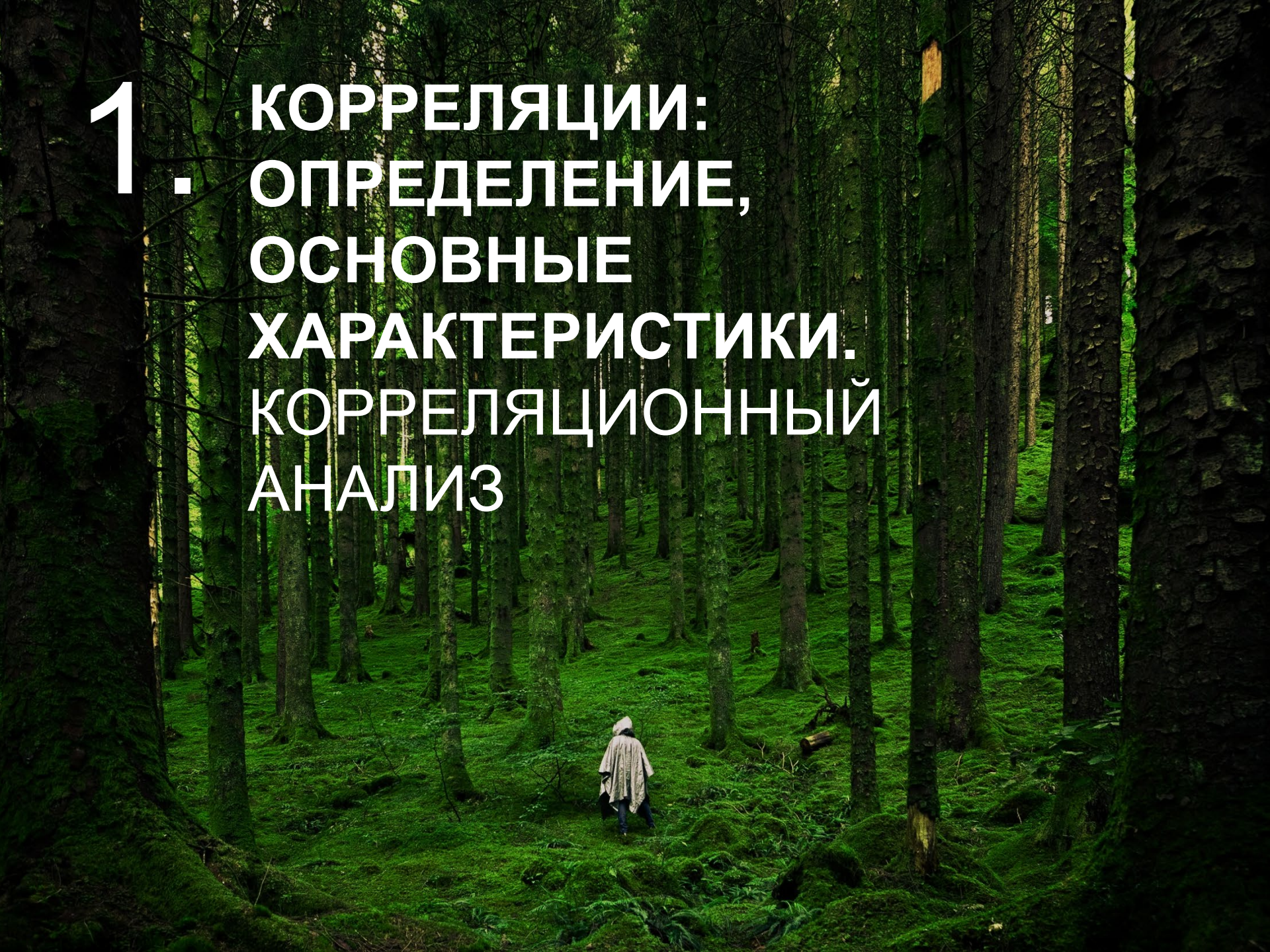
- найти связи между показателями
- проверить исследовательскую гипотезу
- понять направление связи
- оценить силу связи
- подготовить основу для дальнейшего анализа: регрессии, сегментации, моделирования.

Полезный совет

Коэффициенты корреляции можно использовать для расчета различных индексов.

Ознакомьтесь с вариантами расчета индексов на странице «Индексы и рейтинги»





1. КОРРЕЛЯЦИИ: ОПРЕДЕЛЕНИЕ, ОСНОВНЫЕ ХАРАКТЕРИСТИКИ. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

1. Корреляционный анализ

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

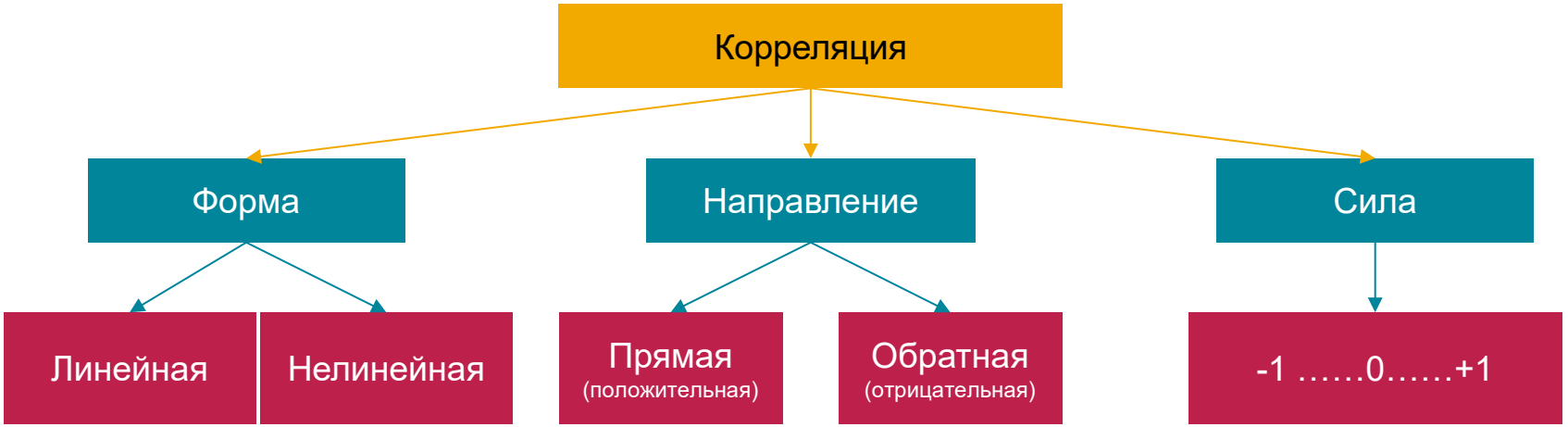
Корреляционная зависимость – это согласованные изменения двух (парная корреляционная связь) или большего количества признаков (множественная корреляционная связь). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой(-их)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

Виды связи между переменными

1. **Прямая причинно-следственная связь** - переменная X определяет значение переменной У.
   Пример: Наличие воды ускоряет рост растений. Яд вызывает смерть. Температура воздуха прямо влияет на скорость таяния льда.

2. **Обратная причинно-следственная связь** - переменная У 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?

Корреляционный анализ сам по себе не устанавливает причинность, а только фиксирует совместную изменчивость.

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

Стаж менеджера по продажам на фирме

Годовой объем продаж

Множественная связь

Успеваемость студента

Успеваемость в школе

Коэффициент IQ

Время на занятия

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.

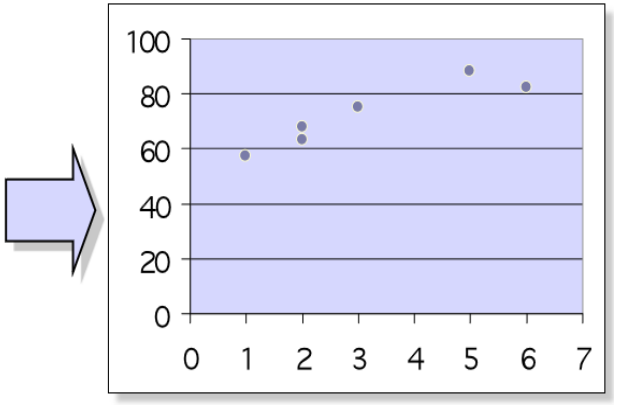
- Команда **«Графика»** → **«Рассеяния/Точки».**



График рассеяния (Scatter Plot)

Пример: Рассматриваем две переменные: «Продолжительность подготовки (часов)» студентов перед экзаменом и «Итоговая оценка» (из 100 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 (при r=0 отсутствует линейная связь; при этом может существовать нелинейная связь)

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

1. Корреляционный анализ

Сила корреляции

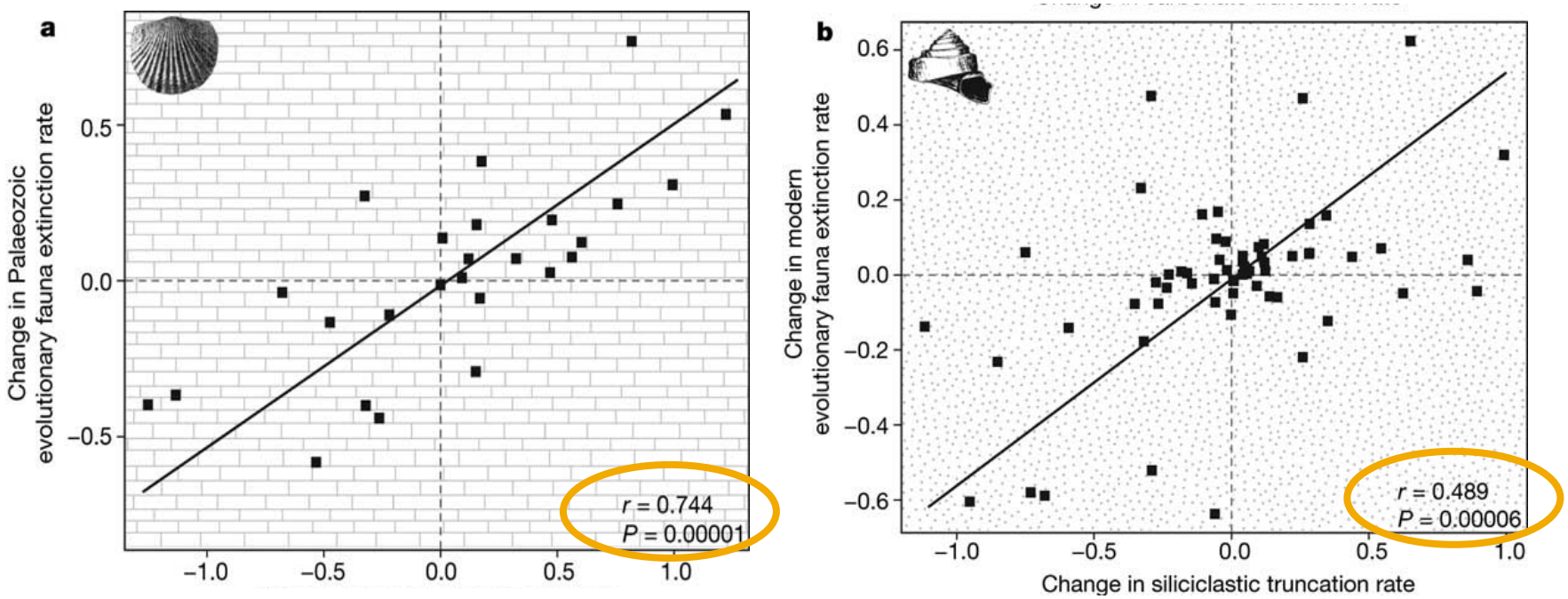
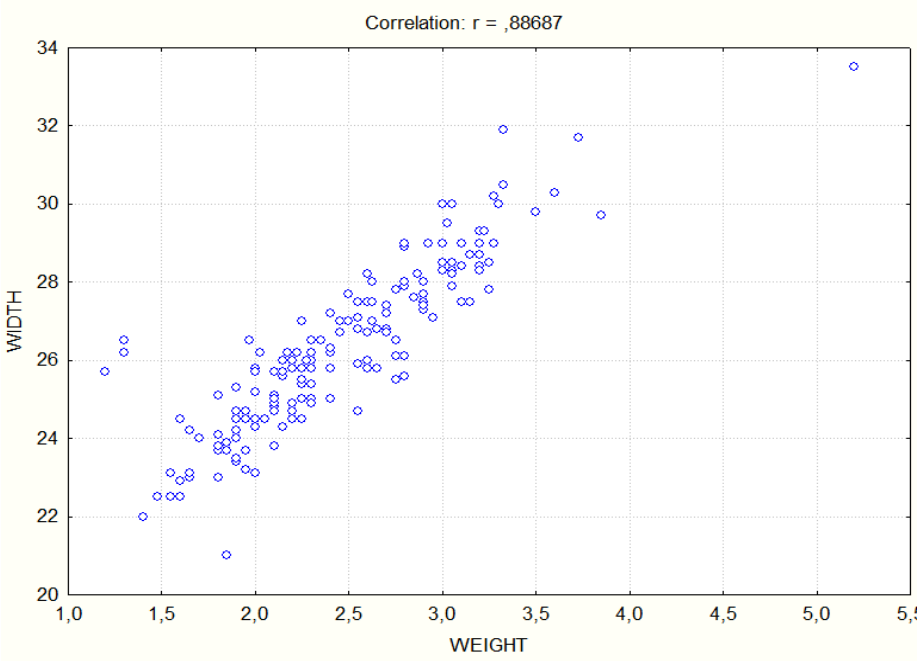


Диаграмма рассеяния (Scatterplot, Scatter diagram)

Характеристики диаграммы:

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

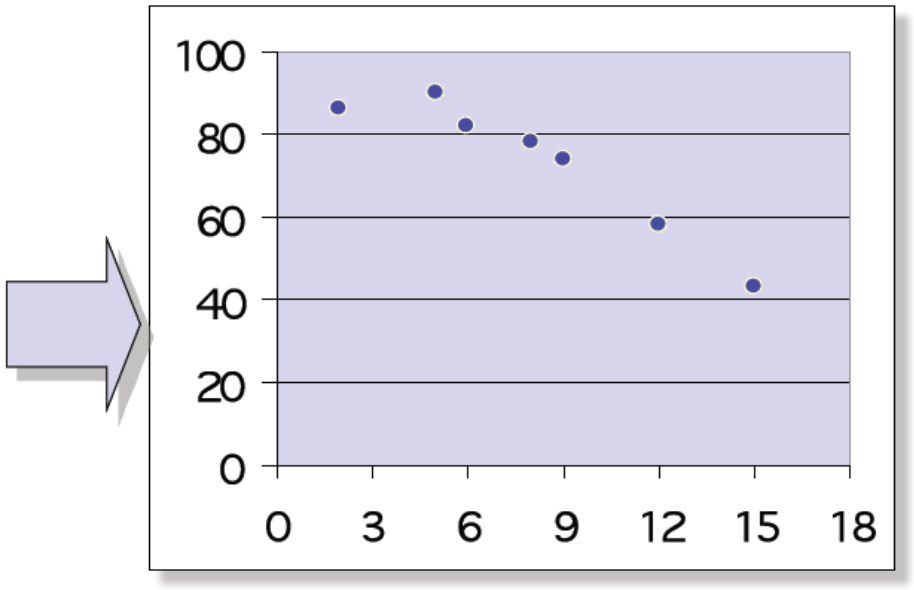
О силе связи можно судить по тому, насколько тесно расположены точки-объекты около линии регрессии - чем ближе точки к линии, тем сильнее связь.



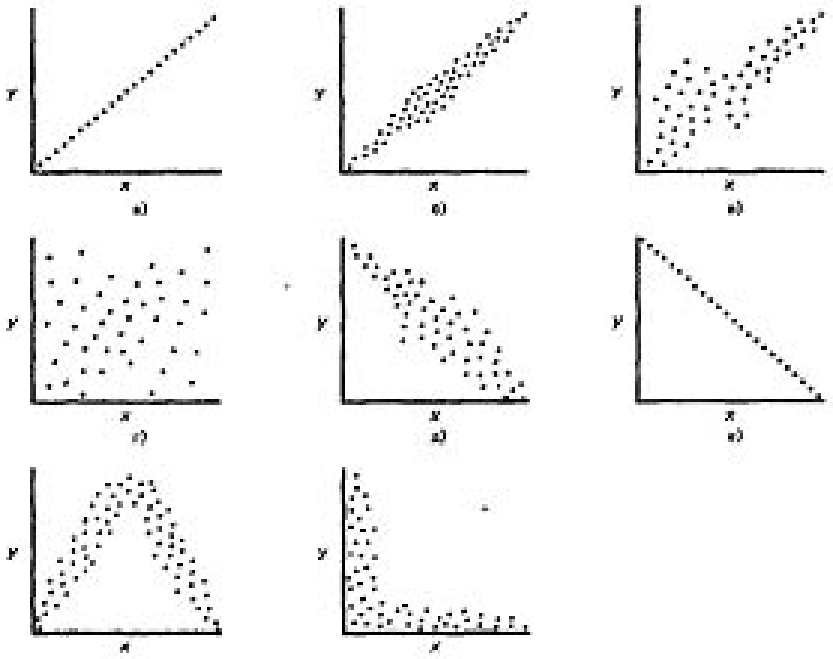
Направление корреляции

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»**.

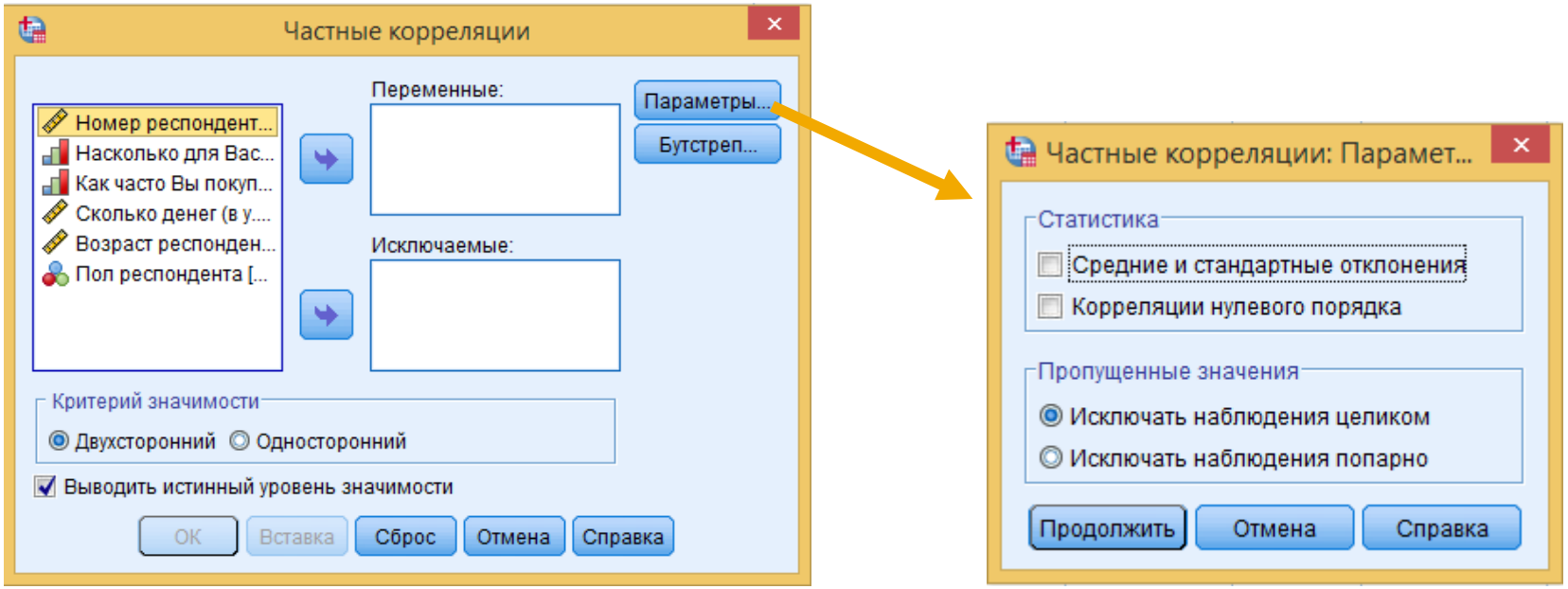
  Пример: «Аисты приносят детей»
  Изучалась корреляция между числом аистов, свивших гнезда в южных районах Швеции, и рождаемостью в эти же годы в Швеции. Вычисления показали высокую положительную корреляцию между этими явлениями. Однако причинная зависимость не может быть выведена ни из какого наблюдаемого 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**.

Частная корреляция



- Перенести необходимые переменные для вычисления корреляции в «Переменные».
- Перенести дополнительную переменную, которая предположительно влияет на вышеуказанные переменные, в «Исключаемые».

Коэффициенты корреляции

1. Для порядковых данных используются следующие **коэффициенты корреляции**:
   - ρ - коэффициент ранговой корреляции Спирмена
   - τ - коэффициент ранговой корреляции Кендалла
   - γ - коэффициент ранговой корреляции Гудмена – Краскела
2. Для переменных с интервальной и номинальной шкалой используется **коэффициент корреляции Пирсона**.
3. Если, по меньшей мере, одна из двух переменных имеет порядковую шкалу, либо не является нормально распределённой, используется ранговая корреляция **Спирмана** или **т-Кендалла**. Применение коэффициента **Кендалла** предпочтительно, если в исходных данных имеются выбросы.

Типы шкал		Мера связи
Переменная X	Переменная Y	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла
Дихотомическая	Дихотомическая	Коэффициент φ,
Дихотомическая	Ранговая	Рангово-бисериальный коэффициент
Дихотомическая	Интервальная или отношений	Бисериальный коэффициент
Интервальная	Ранговая	Не разработан

Взаимосвязь количественных переменных

Пример:

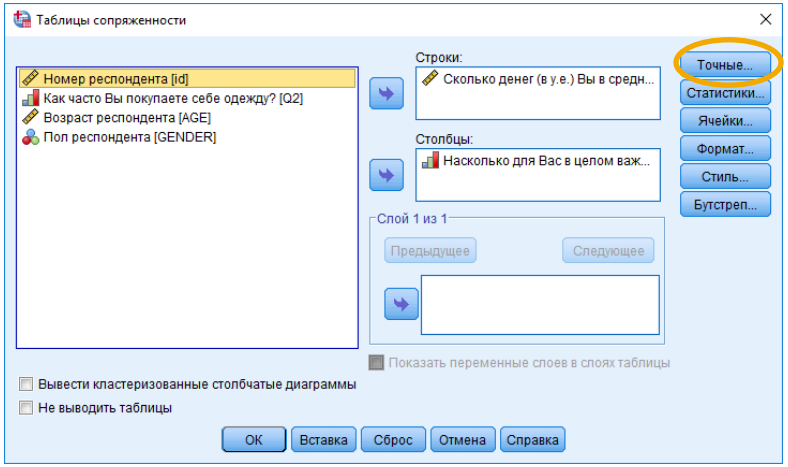
1. Массив данных fashion.sav.
2. Задача: Узнать, есть ли зависимость между интересом к моде (Q1) и тем, сколько денег человек тратит за один поход в магазин за одеждой (Q3).

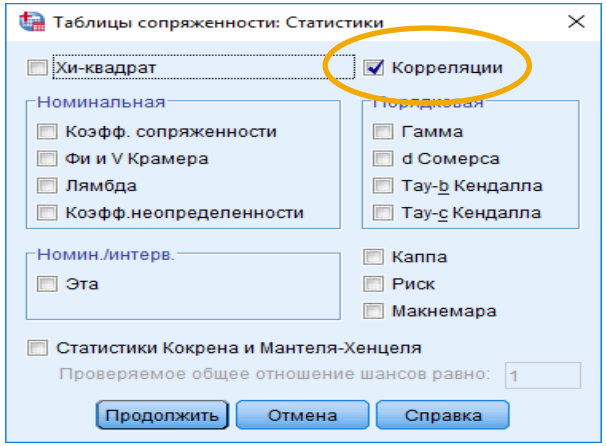
Взаимосвязь количественных переменных

Варианты расчета коэффициентов

Вариант 1

Команда **«Таблицы сопряженности» → «Статистики» → «Корреляции»**
(рассчитываются коэффициенты Пирсона и Спирмена для двух переменных)





Симметричные меры

		Значение	Асимптотическая среднеквадратичная ошибка[a]	Примерная T[b]	Примерная Знач.
Интервал/интервал	R Пирсона	,259	,051	3,781	,000[c]
Порядковый/порядковый	Корреляция Спирмена	,344	,066	5,148	,000[c]
Количество допустимых наблюдений		200			

a. Не предполагая нулевой гипотезы.

b. Использование асимптотической среднеквадратичной ошибки в предположении нулевой гипотезы.

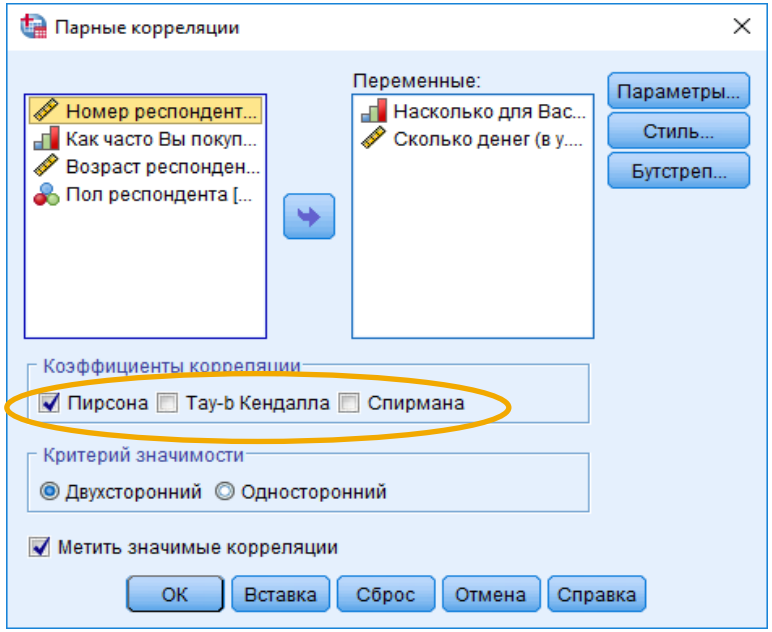
c. Основано на нормальной аппроксимации.

1. Корреляционный анализ

Взаимосвязь количественных переменных

Вариант 2

Команда **«Анализ»** → **«Корреляции»** → **«Парная»**
(рассчитываются коэффициенты Пирсона, Спирмена, Кендалла попарно для любого количества переменных)



Корреляции

		Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?	Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?
Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?	Корреляция Пирсона	1	,259**
	Знач. (двухсторонняя)		,000
	N	200	200
Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	Корреляция Пирсона	,259**	1
	Знач. (двухсторонняя)	,000	
	N	200	200

**. Корреляция значима на уровне 0,01 (двухсторонняя).

Корреляции

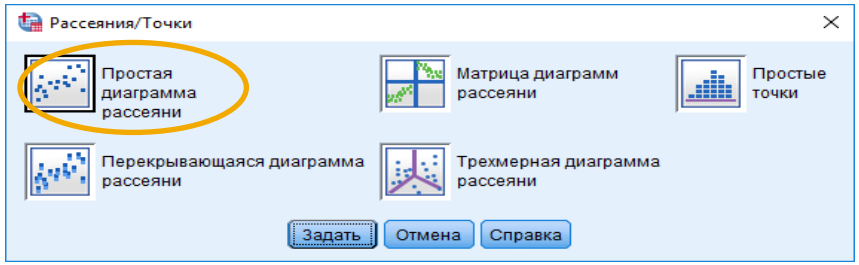
			Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?	Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?
Ро Спирмана	Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?	Коэффициент корреляции	1,000	,344**
		Знач. (2-х сторонняя)	.	,000
		N	200	200
	Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	Коэффициент корреляции	,344**	1,000
		Знач. (2-х сторонняя)	,000	.
		N	200	200

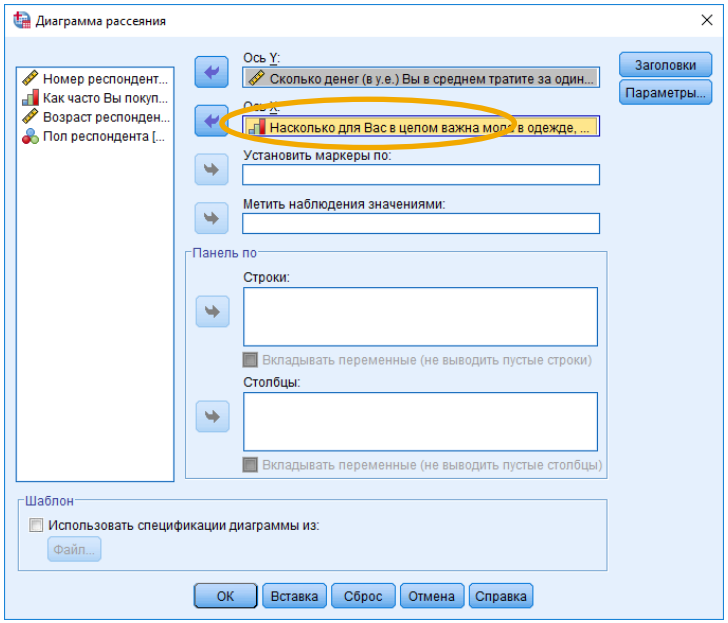
**. Корреляция значима на уровне 0,01 (двухсторонняя).

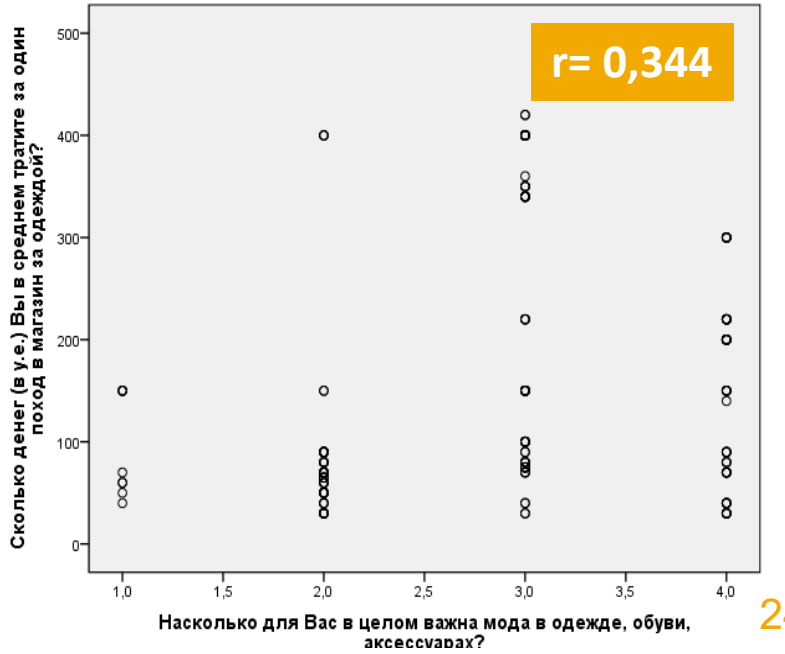
Взаимосвязь количественных переменных

Вариант 3

Команда **«Graphs»** → **«Scatter»** → **«Simple»** (графическая визуализация)



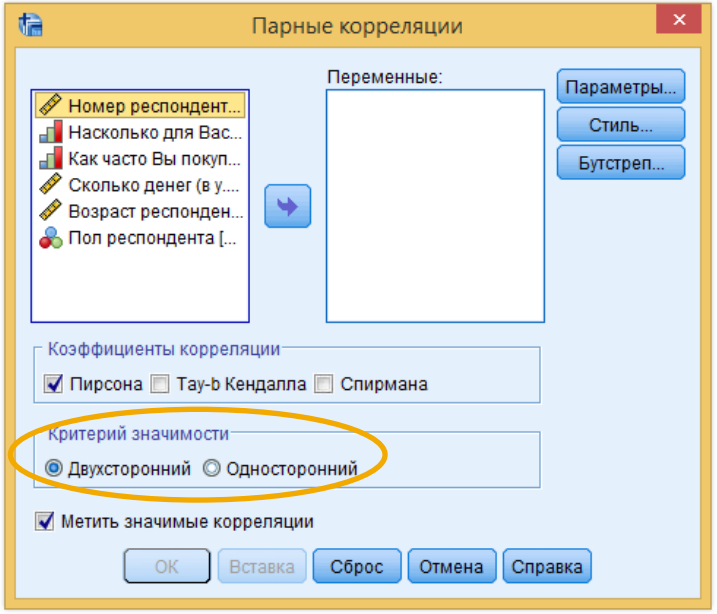




Уровень значимости

Помимо значений корреляции, вычисляются уровни значимости (p-value). В SPSS можно использовать односторонний и двусторонний тест значимости.

Обычно используют **двусторонний.**

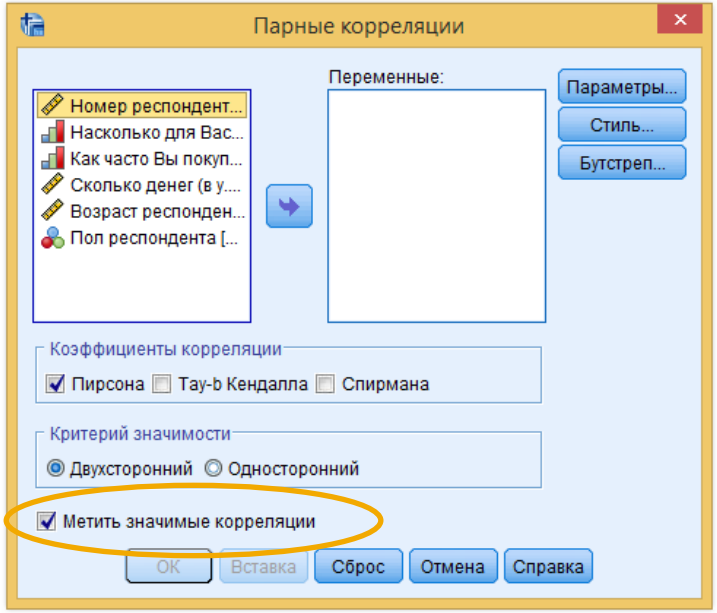


Значимость (p-value) показывает, есть ли основания считать корреляцию отличной от нуля в генеральной совокупности.

Уровень значимости

По умолчанию отмечено «Метить значимые корреляции»

Уровень значимости	Помечены в SPSS значения корреляции
От 0,01 до 0,05	*
От 0 до 0,01	**





2. ПАРНЫЕ СТАТИСТИЧЕСКИЕ СВЯЗИ



2.1 КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА (МЕТРИЧЕСКИЕ ШКАЛЫ)

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.

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

стандартное отклонение для веса

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

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 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
	$\Sigma x=19$	**$\Sigma y=433$**	**$\Sigma xy=1476$**	**$\Sigma x^2=79$**	**$\Sigma y^2=31935$**

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

Оценка статистической значимости коэффициента корреляции

Критическое значение *t*-критерия определяется из таблицы значений *t*-распределения для выбранного уровня значимости 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 ХИ-квадрат.



2.2 КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА (НЕМЕТРИЧЕСКИЕ ШКАЛЫ)

2.2. Коэффициент ранговой корреляции Спирмена

Коэффициенты ранговой корреляции используются для измерения взаимозависимости между качественными признаками, значения которых могут быть упорядочены или проранжированы по степени убывания (или возрастания) данного качества у исследуемых социальных объектов.

Метод ранговой корреляции Спирмена позволяет определить тесноту (силу) и направление корреляционной связи между двумя признаками или двумя профилями (иерархиями) признаков.

- Для подсчета ранговой корреляции необходимо располагать двумя рядами значений, которые могут быть проранжированы.
- Возможны два варианта гипотез коэффициента ранговой корреляции Спирмена:

Но: Корреляция между переменными А и Б не отличается от нуля.
Н1: Корреляция между переменными А и Б достоверно отличается от нуля.

Но: Корреляция между иерархиями А и Б не отличается от нуля.
Н1: Корреляция между иерархиями А и Б достоверно отличается от нуля.

2.2. Коэффициент ранговой корреляции Спирмена

Коэффициент ρ Спирмена интерпретируется аналогично коэффициенту корреляции Пирсона и может принимать значения в таком же диапазоне (–1;+1).

$$r_s = 1 - \frac{6\sum_{i=1}^{l} d_i^2}{l\left(l^2 - 1\right)}$$

Где:

$\sum_{i=1}^{l} d_i^2$ – сумма квадратов разностей рангов

l – число парных наблюдений



2.3 КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ КЕНДАЛЛА

2.3. Коэффициент ранговой корреляции Кендалла

- Как и коэффициент ранговой корреляции Спирмена, **коэффициент Кендалла** используется для измерения взаимосвязи между качественными признаками, характеризующими объекты одной и той же природы, ранжированные по одному и тому же критерию. Изменяется от -1 до +1.

$$\tau_a = \frac{S}{1/2l(l-1)}$$

- Коэффициенты корреляции Спирмена и Кендалла используются как меры взаимозависимости между **рядами рангов**, а не как меры связи между самими переменными.

- Коэффициенты Спирмена и Кендалла обладают примерно одинаковыми свойствами, но тау-Кендалла в случае многих рангов, а также при введении дополнительных объектов в ходе исследования имеет определенные вычислительные преимущества.

2.3. Коэффициент ранговой корреляции Кендалла

Пример: Оценивается связь между ростом и весом в группе людей, предварительно ранжированных по этим переменным.

При сравнении любых двух человек из этой группы возможны две ситуации:

- однонаправленное изменение переменных **(«совпадение»),** когда и рост, и вес одного больше, чем у другого;
- разнонаправленное изменение **(«инверсия»),** когда рост у второго больше, а вес меньше, чем у первого.

2.3. Коэффициент ранговой корреляции Кендалла

Перебрав все пары испытуемых, можно оценить вероятность совпадений (P) и вероятность инверсий (Q). **Корреляция Кендалла** — это разность вероятностей «совпадений» и «инверсий»:

$$\tau = P - Q$$

По значению корреляции Кендалла можно всегда вычислить вероятность «совпадений» и «инверсий» :

$$P = (1 + \tau)/2)$$

$$(Q = (1 - \tau)/2)$$

Важным преимуществом корреляции **т-**Кендалла является ее отчетливая вероятностная интерпретация.



2.4 КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ ГУДМЕНА-КРАСКЕЛА

2.4. Коэффициент ранговой корреляции Гудмена - Краскела

- Альтернативой мерам ассоциации, построенным на ХИ-квадрат статистике, являются коэффициенты, основанные на идее Л. Гудмана и Е. Краскала о «**пропорциональной редукции ошибок**» (Proportional Reduction in Error Measures — PRE).

- В отличие от предыдущих показателей коэффициент PRE предполагает четкое разграничение на зависимые (Y) и независимые (X) переменные. Например, в качестве независимой переменной может выступать строковая переменная «уровень дохода», а в качестве зависимой — столбцовая переменная «степень удовлетворенности».

- В статистическом распределении двух переменных содержится определенная информация об их зависимости. Если переменная X влияет на переменную Y, то, зная распределение независимой переменной, можно сделать вывод о характере распределения зависимой переменной. Естественно, что эта оценка не всегда правильна. Существует вероятность ошибки.

2.4. Коэффициент ранговой корреляции Гудмена - Краскела

- В случае **полной зависимости** двух переменных друг от друга на основе информации о значении независимой переменной для каждой единицы наблюдения можно совершенно точно ”предсказать значение“ зависимой.
- При **полной независимости** переменных это сделать не удается.
- Таким образом, по тому, как увеличивается точность прогноза значения зависимой переменной с учетом дополнительной информации о независимой переменной, определяют степень их зависимости.
- Подобные показатели можно формировать для **любых типов шкал**.

Литература по Теме 6

1. **Бююль А., Цеффель П. SPSS: искусство обработки информации. – М., 2005**
   - Глава 15. Корреляции
2. **Наследов А. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. – СПб., 2013**
   - Глава 9. Корреляции
3. **Сибирев В.А. «Введение в анализ социальных данных» (С. 58-81)**

Памятка «Как выбрать коэффициент корреляции?»

Данные	Что использовать
Две количественные переменные, линейная связь	Пирсон
Порядковые шкалы / ранги / монотонная связь	Спирмен
Порядковые шкалы, много совпадающих рангов / небольшие выборки	Кендалл
Две номинальные переменные	Cramer’s V, Phi
Нужно учесть третью переменную	Частная корреляция

Чек-лист аналитика

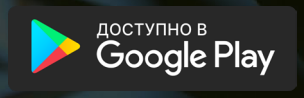
- ❑ построена диаграмма рассеяния;
- ❑ выбран корректный коэффициент;
- ❑ проверены выбросы;
- ❑ p-value интерпретирован правильно;
- ❑ сила связи не преувеличена;
- ❑ причинность не заявлена без дополнительных оснований;
- ❑ проверены возможные третьи факторы.

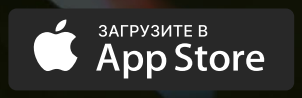
Корреляционный анализ показывает, насколько согласованно изменяются две переменные. Он помогает найти направление и силу связи, но не доказывает причинность. Для корректного вывода важно выбрать коэффициент по типу шкалы, проверить значимость, посмотреть график и подумать о возможных третьих факторах.

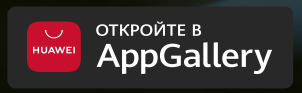
НАФИ

О том, как на практике применяется кластерный анализ вы можете увидеть, изучив, например, страницы решений «Исследования потребительского поведения (U&A)» или «Исследование показателей «здоровья» бренда (BHT)» на нашем сайте.

Хотите узнать, как проводятся онлайн-опросы, и попробовать себя в роли респондента? Регистрируйтесь на нашей платформе https://tetotvet.ru или скачивайте наше приложение «Тет-о-твет» в любом из магазинов приложений.







Для свободного использования в образовательных целях
Copyright 2017 © Академия НАФИ. Москва
Все права защищены
nafi.ru