Лекция 2

Корреляционный и регрессионный анализ

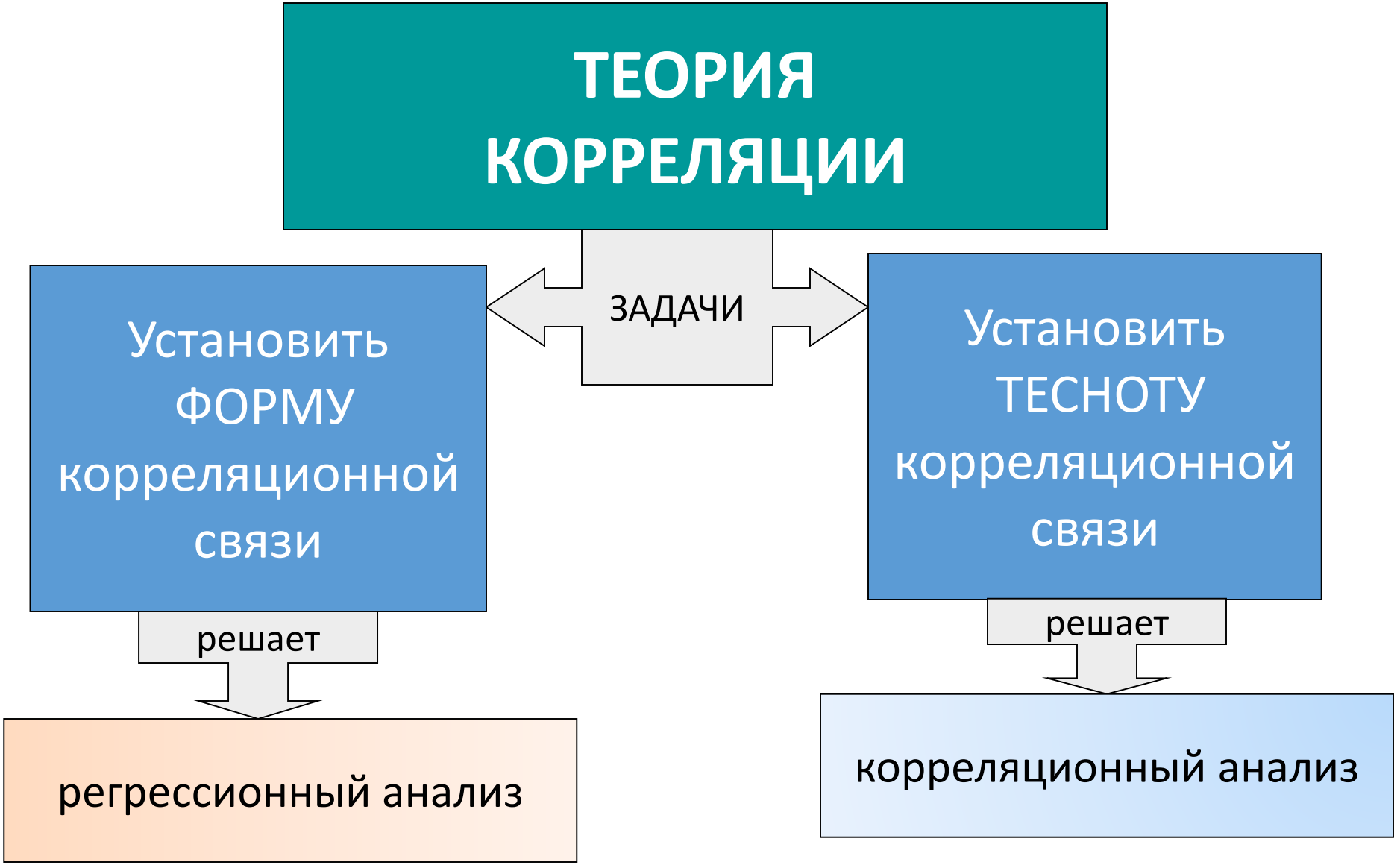
Понятие корреляционной зависимости

Многие задачи требуют установить и оценить зависимость между двумя или несколькими случайными величинами.

- Зависимость случайных величин называют ***статистической***, если изменение одной величины влечет изменение распределения другой величины.
- Статистическая зависимость называется ***корреляционной***, если при изменении одной величины изменяется среднее значение другой.

- Если случайная величина представляет некоторый признак (например, статистические наблюдения некой экономической величины), то под **корреляцией** понимают – меру согласованности одного признака с другим, или с несколькими, либо взаимную согласованность группы признаков.

- Функциональная зависимость предполагает взаимно однозначное соответствие аргумента x и функции $y=f(x)$, вероятностная же зависимость допускает некий условный диапазон, в который предположительно (с такой-то долей вероятности) попадает значение признака y_i при значении x_i признака x.


ТЕОРИЯ КОРРЕЛЯЦИИ
ЗАДАЧИ
Установить ФОРМУ корреляционной связи
Установить ТЕСНОТУ корреляционной связи
решает
решает
регрессионный анализ
корреляционный анализ

Корреляционный анализ

- **Корреляционный анализ** – один из методов исследования взаимосвязи между двумя или более переменными.
- Для применения линейного корреляционного анализа величины, образующие пары, должны быть распределены нормально.
- Корреляционная зависимость характеризуется ***формой и теснотой связи***.
- **Функция регрессии** определяет форму связи при изучении статистических зависимостей, а тесноту связи определяют с помощью коэффициента корреляции.

Корреляционный анализ

- **В качестве числовой характеристики вероятностной связи используют коэффициенты корреляции**, значения которых изменяются в диапазоне от –1 до +1. После проведения расчетов исследователь, как правило, отбирает только наиболее сильные корреляции, которые в дальнейшем интерпретируются
- ***Критерием для отбора «достаточно сильных» корреляций*** может быть как абсолютное значение самого коэффициента корреляции (от 0,7 до 1), так и относительная величина этого коэффициента, определяемая по уровню статистической значимости (от 0,01 до 0,1), зависящему от размера выборки.
- ***В малых выборках*** для дальнейшей интерпретации корректнее отбирать сильные корреляции на основании уровня статистической значимости.
- Для исследований, которые проведены на больших выборках, лучше использовать абсолютные значения коэффициентов корреляции.

Корреляционный анализ. Подготовка данных

- **Измерение** - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.
- В процессе подготовки данных измеряется не сам объект, а его характеристики.
- ***Шкала*** - правило, в соответствии с которым объектам присваиваются числа.
- Переменные могут являться **числовыми** данными либо **символьными**.
- Числовые данные, в свою очередь, могут быть дискретными и непрерывными.
- Существует пять типов шкал измерений: номинальная, порядковая, интервальная, относительная и дихотомическая.

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

Корреляционный анализ. Подготовка данных

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

- 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

Корреляционный анализ. Коэффициенты корреляции

- **В настоящее время разработано множество различных коэффициентов корреляции.** Наиболее применяемыми являются r-Пирсона, r-Спирмена и τ-Кендалла.

- Выбор метода вычисления коэффициента корреляции зависит от типа шкалы, к которой относятся переменные

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

Корреляционный анализ. Подготовка данных. Типы шкал

Название	Содержание	Пример
Номинальная шкала (nominal scale)	шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.	профессии, город проживания, семейное положение
Порядковая (ранговая) шкала (ordinal scale):	шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними.	место (1, 2, 3-е), которое команда получила на соревнованиях, номер студента в рейтинге успеваемости (1-й, 23-й, и т.д.), при этом неизвестно, насколько один студент успешней другого, известен лишь его номер в рейтинге
Интервальная шкала (interval scale):	шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.	температура воды в море утром - 19 градусов, вечером - 24, т.е. вечерняя на 5 градусов выше, но нельзя сказать, что она в 1,26 раз выше
Относительная шкала (ratio scale) или шкала отношений:	шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы. Является числовой	вес новорожденного ребенка (4 кг и 3 кг). Первый в 1,33 раза тяжелее
Дихотомическая шкала (dichotomous scale):	шкала, содержащая только две категории.	пол (мужской и женский)

Корреляционный анализ. Подготовка данных. Типы шкал

1. **Для порядковых данных** используются следующие коэффициенты корреляции:

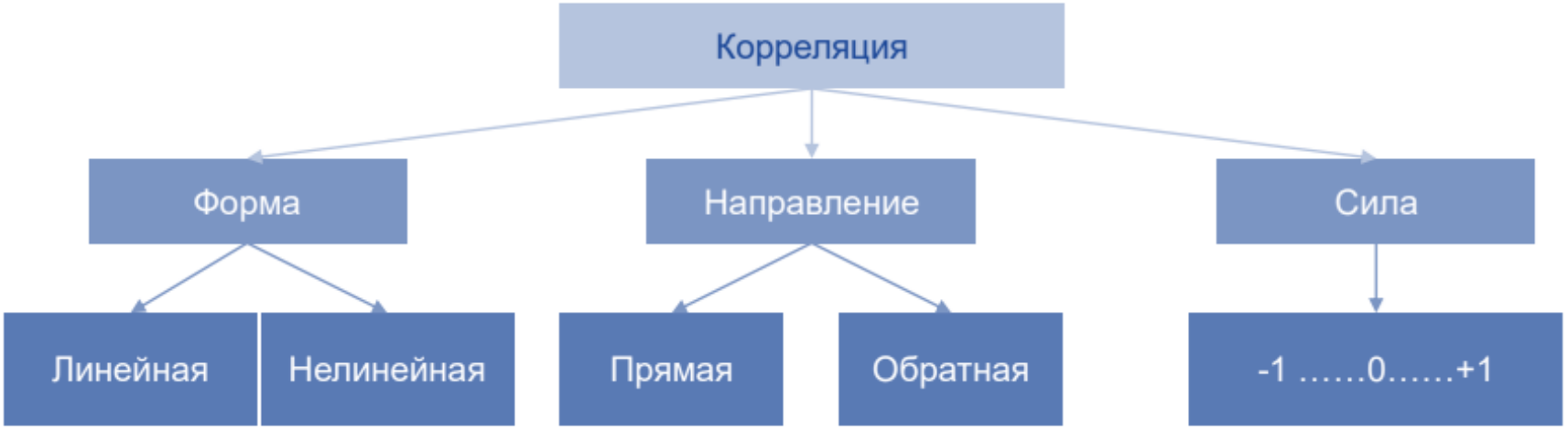
- ρ (r_s)- коэффициент ранговой корреляции Спирмена
- τ - коэффициент ранговой корреляции Кендалла
- γ - коэффициент ранговой корреляции Гудмена – Краскела

2. **Для переменных с интервальной и номинальной шкалой** используется коэффициент корреляции Пирсона (корреляция моментов произведений).

3. **Если, по меньшей мере, одна из двух переменных имеет порядковую шкалу, либо не является нормально распределённой**, используется ранговая корреляция Спирмана или τ-Кендалла.

Применение коэффициента Кендалла предпочтительно, если в исходных данных имеются выбросы.

Корреляционный анализ. Характер связи между переменными



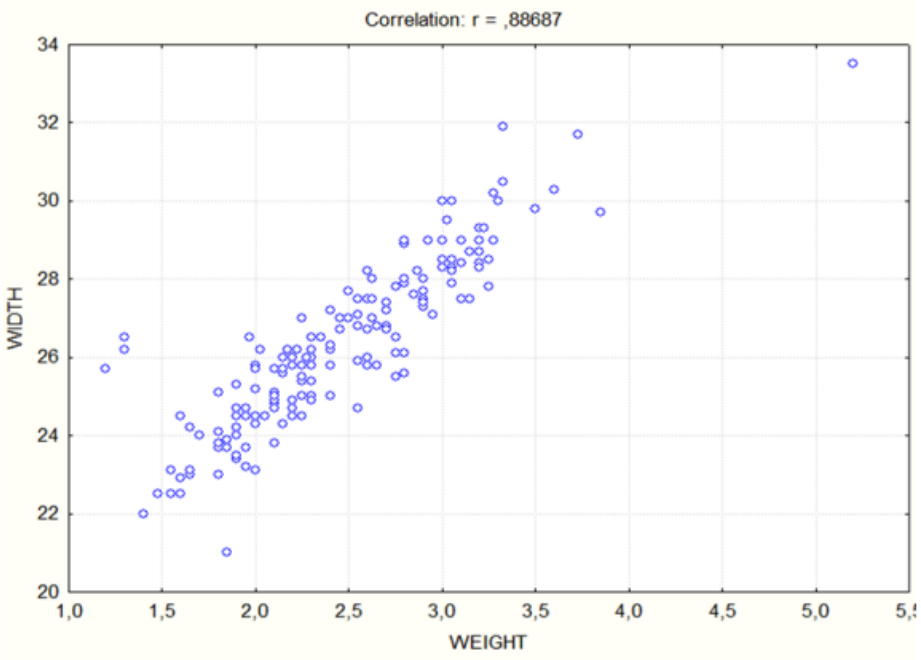
1. Прямая причинно-следственная связь - переменная X определяет значение переменной У.
2. Обратная причинно-следственная связь - переменная У определяет значение переменной X.
3. Связь, вызванная третьей (скрытой) переменной.
4. Связь, вызванная несколькими скрытыми переменными.
5. Связи нет, наблюдаемая зависимость случайна.

Корреляционный анализ. Характер связи между переменными

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.

Коэффициент корреляции Пирсона

- Наиболее часто используемый **коэффициент корреляции Пирсона** r измеряет степень линейных связей между переменными.
- Числовое значение коэффициента корреляции Пирсона определяется формулой:

$$r = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}}.$$

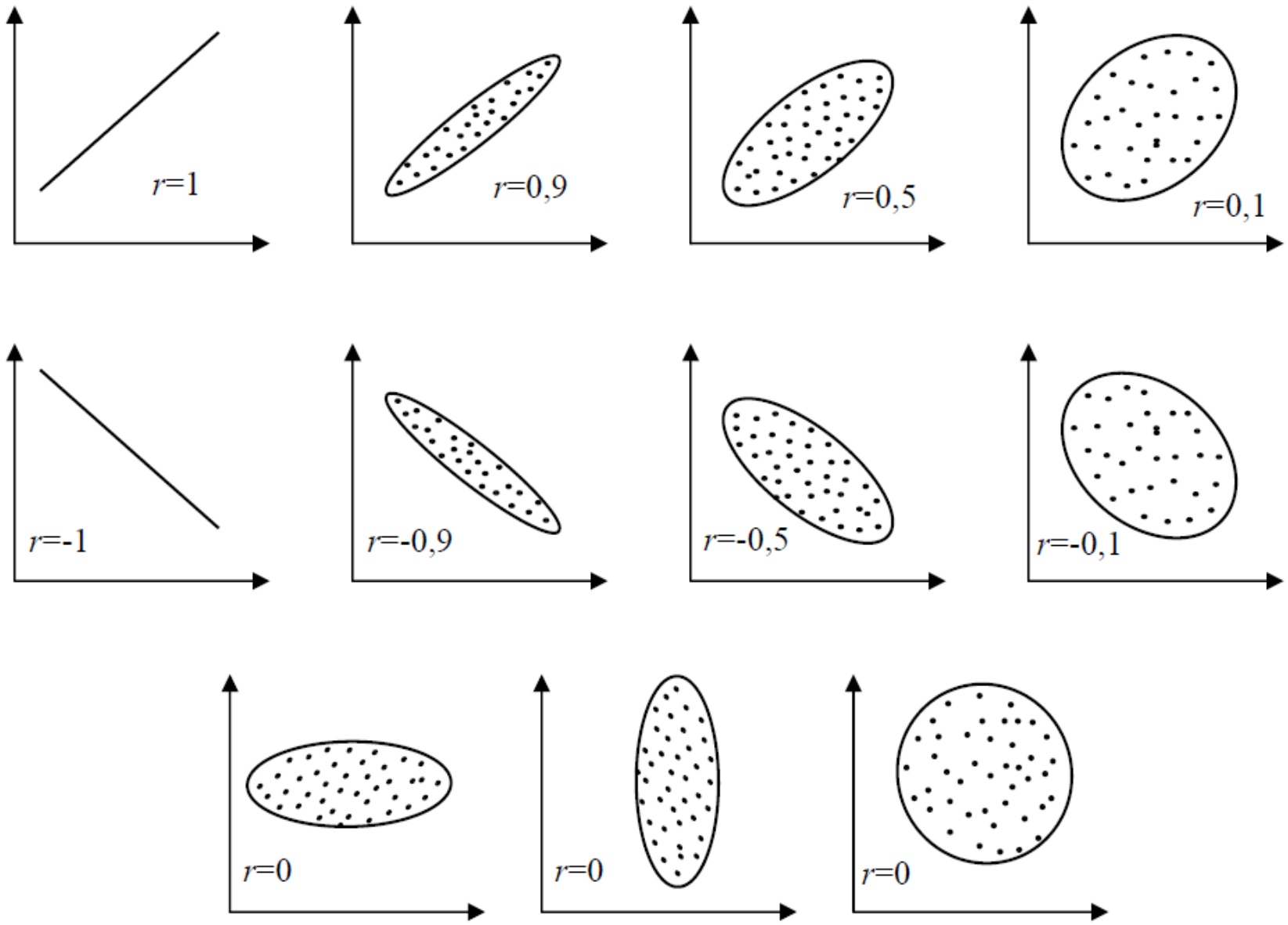
Коэффициент корреляции Пирсона

- Значение коэффициента корреляции r лежит в интервале [-1;1].
- При $r=0$ корреляция отсутствует.
- При $|r|=1$ корреляция является полной или абсолютной.
- Чем ближе $|r|$ к 1, тем теснее связь между переменными.
- Отрицательное значение коэффициента корреляции свидетельствует об обратной зависимости между переменными, положительное значение – о прямой.

Коэффициент корреляции Пирсона

- Чем больше разбросанность точек по всему корреляционному полю, тем слабее зависимость между переменными.
- Если на графике зависимость можно представить прямой линией (с положительным или отрицательным углом наклона), то корреляция между переменными будет высокая.

Примеры корреляционной зависимости



Проверка гипотезы о независимости наблюдений

- Для проверки гипотезы о независимости наблюдений используют **t-критерий Стьюдента**.
- Расчетное значение критерия вычисляется по формуле

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

- Если критическое значение критерия Стьюдента, соответствующее выбранному уровню значимости и числу степеней свободы, равному n - 2 (где n – число пар (x, y)), меньше расчетного ($t_{кр}<t$), то гипотеза о независимости значений X и Y должна быть отвергнута.

Показатель ранговой корреляции Спирмена

- Для определения корреляции порядковых признаков используют **показатель ранговой корреляции Спирмена**.
- Расчет такого коэффициента корреляции не требует нормальности распределения и линейной зависимости от переменных, и он может быть применен как к количественным, так и порядковым признакам.

Показатель ранговой корреляции Спирмена

Идея **коэффициента ранговой корреляции Спирмена** заключается в том, что:

- все данные упорядочиваются по возрастанию переменной, а сами значения заменяются их рангами,
- затем вычисляются разностные ранги, по которым рассчитывается *коэффициент корреляции Спирмена* по формуле

$$r_s = 1 - \frac{6\sum d_i}{n^3 - n},$$

где d_i - разность рангов для каждого члена выборки; n - число пар значений x, y.

Показатель ранговой корреляции Спирмена

Для определения различий между признаками находят критическое значение **коэффициента корреляции Спирмена** для выбранного доверительного уровня и заданного объема выборки. Если объем выборки n>50, то используют критерий Стьюдента

$$t_s = \frac{r_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}} .$$

- Расчетное значение сравнивается с критическим для числа степеней свободы m=n-2 и заданным уровнем значимости.
- Если критическое значение критерия Спирмена меньше, чем расчетное (или критическое значение t-критерия Стьюдента меньше, чем расчетное для случая n>50), то различия считаются статистически значимыми.

Коэффициент корреляции Кенделла

- **Коэффициент ранговой корреляции τ-Кендалла** является самостоятельным оригинальным методом, опирающимся на вычисление соотношения пар значений двух выборок, имеющих одинаковые или отличающиеся тенденции (возрастание или убывание значений).
- Этот коэффициент называют еще *коэффициентом конкордации*.
- ***Основной идеей*** данного метода является то, что о направлении связи можно судить, попарно сравнивая между собой «испытуемых»:
- если у пары «испытуемых» изменение по *X* совпадает по направлению с изменением по *Y*, это свидетельствует о положительной связи,
- если не совпадает – об отрицательной связи, например, при исследовании личностных качеств, имеющих определяющее значение для семейного благополучия.

Коэффициент корреляции Кенделла

В этом методе одна переменная представляется в виде монотонной последовательности в порядке возрастания величин; другой переменной присваиваются соответствующие ранговые места.

Количество инверсий (нарушений монотонности по сравнению с первым рядом) используется в формуле для корреляционных коэффициентов.

- **Коэффициент корреляции τ-Кенделла** (Kendall tau rank correlation coefficient) – мера линейной связи между случайными величинами.
- Корреляция Кенделла является ранговой, то есть для оценки силы связи используются не численные значения, а соответствующие им ранги.
- Коэффициент инвариантен по отношению к любому монотонному преобразованию шкалы измерения.

Коэффициент корреляции Кенделла

- Заданы две выборки $x=(x_1,\ldots,x_n),\ y=(y_1,\ldots,y_n)$

Коэффициент корреляции Кенделла вычисляется по формуле:

$$\tau = 1-\frac{4}{n(n-1)}R$$

где $R=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\left[x_i < x_j\right]\neq\left[y_i < y_j\right]\right]$ - количество инверсий,

образованных величинами y_i расположенными в порядке возрастания соответствующих x_i.

- Коэффициент $\boldsymbol{\tau}$ принимает значения из отрезка [-1, 1].
- Равенство $\boldsymbol{\tau}=1$ указывает на строгую прямую линейную зависимость, $\boldsymbol{\tau}=-1$ на обратную.
- Коэффициент $\boldsymbol{\tau}$ (линейно связанный с R) можно считать мерой неупорядоченности второй последовательности относительно первой.

Коэффициент корреляции Кенделла

- Другая форма записи коэффициента корреляции Кенделла (или *коэффициента корреляции рангов Кендалла*):

$$\tau = \frac{2S}{n(n-1)},$$

где S=P+Q.

Для нахождения суммы S находят два слагаемых P и Q.

- При определении слагаемого P нужно установить, сколько чисел, находящихся справа от каждого из элементов последовательности рангов переменной y, имеют величину ранга, превышающую ранг рассматриваемого элемента.
- Второе слагаемое Q характеризует степень несоответствия последовательности рангов переменной y последовательности рангов переменной x.
- Чтобы определить Q подсчитаем, сколько чисел, находящихся справа от каждого из членов последовательности рангов переменной y имеет ранг меньше, чем эта единица. Такие величины берутся со знаком минус.

При достаточно большом числе наблюдений между коэффициентами корреляции рангов Спирмена и коэффициентом корреляции рангов Кендалла существует следующее соотношение: $\rho = \frac{3}{2}\tau$.

Коэффициент сопряженности Бравайса

- В случае, если данные представлены в номинальной шкале типа «*да*» и «*нет*» (т.е. имеется таблица сопряженности признаков 2х2), то для выяснения тесноты связи используется специальная форма коэффициента корреляции Пирсона, которая носит название **коэффициента сопряженности Бравайса** (*или **коэффициента контингенции Пирсона***).

Расчет коэффициента сопряженности Бравайса проводится по формуле

$$C = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$$

где a, b, c, d - значения в клетках таблицы 2х2, а признаки – альтернативные.

Сопряженность признаков

1-й признак	2-й признак		Всего
	да	нет	
Да	a	b	$a+b$
Нет	c	d	$c+d$
Итого	$a+c$	$b+d$	$n=a+b+c+d$

Коэффициент ассоциации

- При исследовании степени тесноты связи между качественными признаками, каждый из которых представлен в виде альтернативного признака, также часто используют **коэффициент ассоциации**

$$K_A = \frac{ad - bc}{ad + bc}$$

- Однако , в тех случаях, **когда хотя бы один из четырех показателей в таблице «четырех полей» отсутствует,** величина коэффициента ассоциации будет равна единице, что дает преувеличенную оценку степени тесноты связи между признаками, и **предпочтение следует отдать коэффициенту контингенции *C*.**

Пример. Нужно оценить влияют ли существующие формы повышения квалификации преподавателей университета на уровень их профессионального мастерства, располагая данными о результатах аттестации студентами 320 преподавателей, из которых 240 повысили квалификацию (данные заносят в таблицу сопряженности).

Коэффициент ассоциации

Пример. Нужно оценить влияют ли существующие формы повышения квалификации преподавателей университета на уровень их профессионального мастерства, располагая данными о результатах аттестации студентами 320 преподавателей, из которых 240 повысили квалификацию (данные заносят в таблицу сопряженности).

Группы преподавателей	Средний балл по сравнению с предыдущим по результатам аттестации		Всего
	не изменился и возрос	снизился	
Повысившие квалификацию по одной из принятых форм	163 (a)	77 (b)	240
Не прошедшие повышение квалификации по принятым формам	46 (c)	34 (d)	80
Всего	209	111	320

$$K_A = \frac{ad - bc}{ad + bc} = \frac{163 \cdot 34 - 77 \cdot 46}{163 \cdot 34 + 77 \cdot 46} = \frac{2000}{9084} = 0{,}22 .$$

$$C = \frac{163 \cdot 34 - 77 \cdot 46}{\sqrt{(163 + 77)(163 + 45)(34 + 77)(34 + 46)}} = 0{,}0947.$$

Вывод: по результатам проведенного в университете обследования вряд ли можно сделать убедительный вывод о повышении профессионального мастерства преподавателей в связи с повышением квалификации по одной из принятых форм (стажировка, факультет повышения квалификации, творческий отпуск и др.), поскольку степень тесноты связи невелика.

- Рассмотрим некоторые свойства полученных оценок b_0 и b_1:
- 1) b_0, b_1 независимые случайные величины;
- 2) оценки подчиняются нормальному закону распределения;
- 3) математическое ожидание $M(b_0)=a_0$ и $M(b_1)=a_1$;
- 4) ковариация $cov(b_0,b_1)=0$;
- 5) дисперсии оценок равны

$$\sigma_{b_0}^2 = \sigma^2 / n, \qquad \sigma_{b_1}^2 = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

где σ^2 – дисперсия ошибок.