Лекция 2

Корреляционный и регрессионный анализ

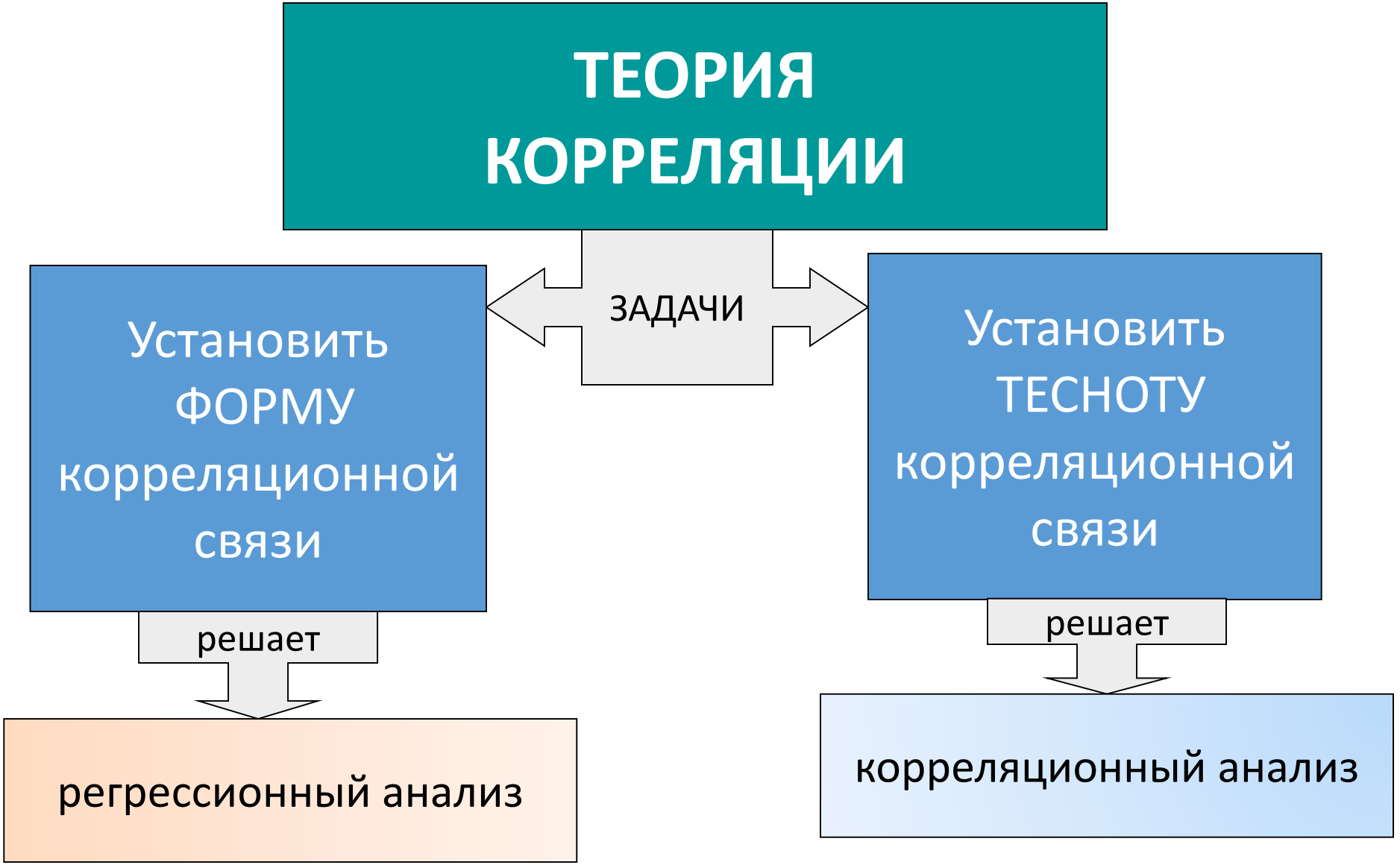
Понятие корреляционной зависимости

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:

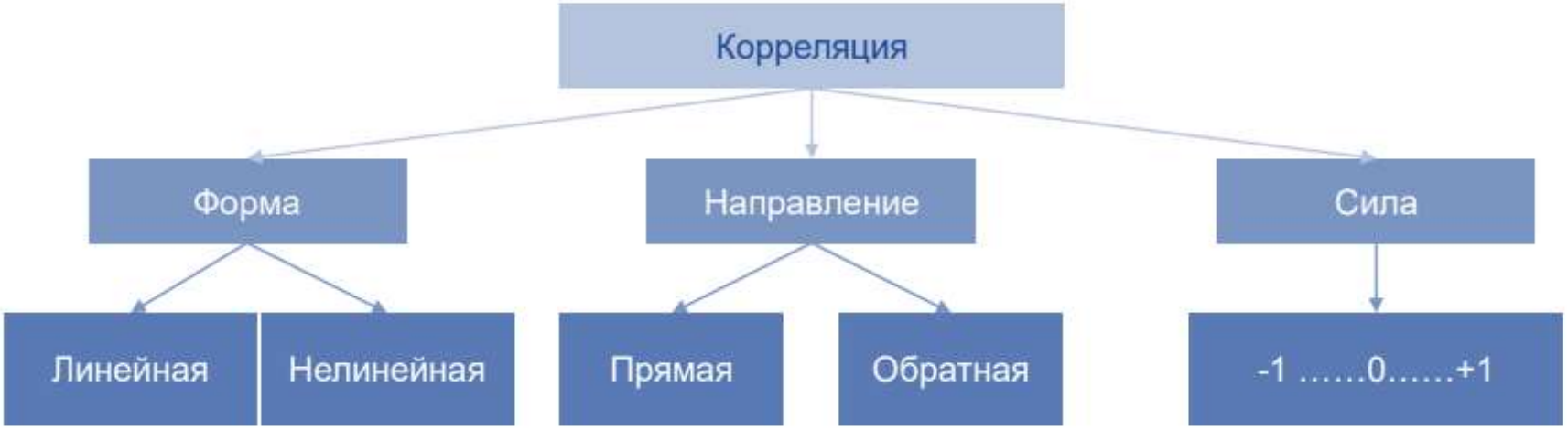
- ρ (r_s)- коэффициент ранговой корреляции Спирмена
- τ - коэффициент ранговой корреляции Кендалла
- γ - коэффициент ранговой корреляции Гудмена – Краскела

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



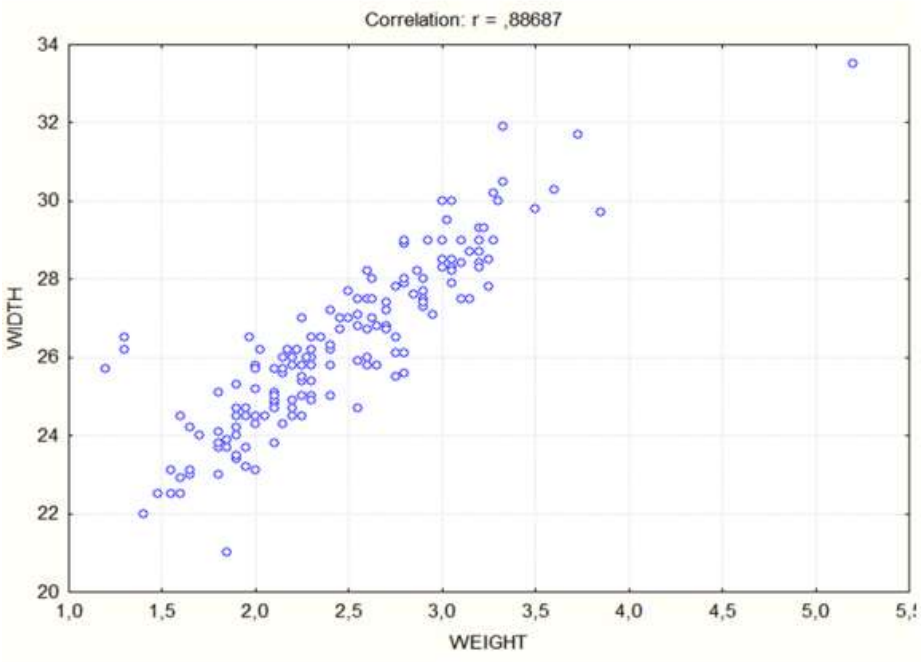
1. Прямая причинно-следственная связь - переменная X определяет значение переменной У.
2. Обратная причинно-следственная связь - переменная У определяет значение переменной X.
3. Связь, вызванная третьей (скрытой) переменной.
4. Связь, вызванная несколькими скрытыми переменными.
5. Связи нет, наблюдаемая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