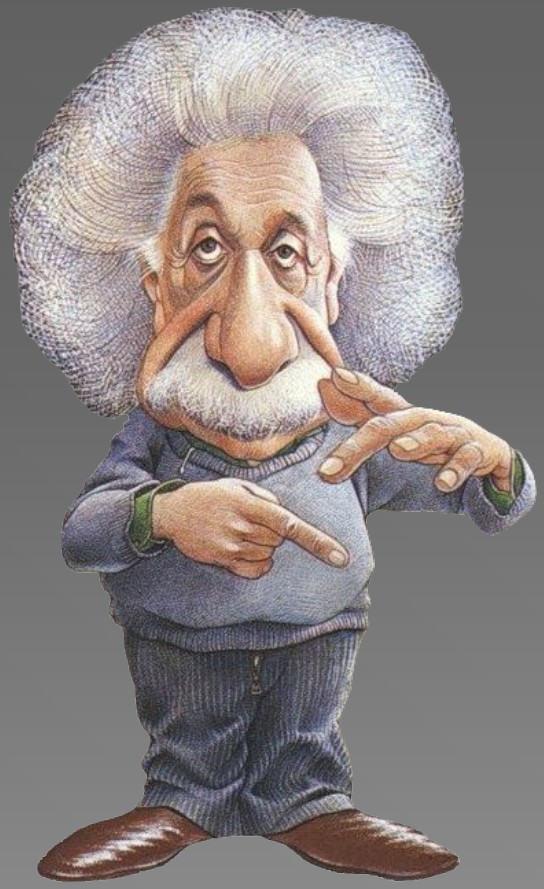
Тема 12

Взаимосвязи между социально-экономическими явлениями:

корреляционный и регрессионный анализ



Вопросы

- Причинность, регрессия, корреляция
- Задачи и предпосылки применения корреляционно-регрессионного анализа
- Парная регрессия на основе метода наименьших квадратов и метода группировок
- Множественная (многофакторная) регрессия
- Оценка значимости параметров взаимосвязи

Причинно-следственные отношения

- - Это связь явлений и процессов, при которой изменение одного из них - причины - ведет к изменению другого - следствия.
- ***Причина*** - это совокупность условий, обстоятельств, действие которых приводит к появлению следствия

Причинно-следственные отношения

- Признаки, обуславливающие изменения других, связанных с ними признаков, называются ***факторными***, или ***факторами***.
- Признаки, изменяющиеся под действием факторных признаков, *являются* ***результативными***.
- Связи между явлениями и их признаками классифицируются по ***степени тесноты***, ***направлению*** и ***аналитическому выражению***.

Причинно-следственные отношения

- Если причинная зависимость проявляется в общем, среднем при большом числе наблюдений, то зависимость называется ***стохастической***.
- Частный случай стохастической зависимости - ***корреляционная*** связь

Формы проявления взаимосвязей

Функциональная связь **(полная)**

- величине факторного признака **строго** соответствует одно или несколько значений функции

Корреляционная **связь (неполная)**

- проявляется в **среднем**, для массовых наблюдений, когда значениям зависимой переменной соответствует ряд **вероятных** значений

Корреляционная связь

- -существует когда изменение среднего значения результативного признака обусловлено изменением факторных признаков.

При корреляционной связи:

Связь между признаками проявляется лишь в среднем, в массе случаев.

Каждому значению аргумента соответствуют случайно распределенные в некотором интервале значения функции.

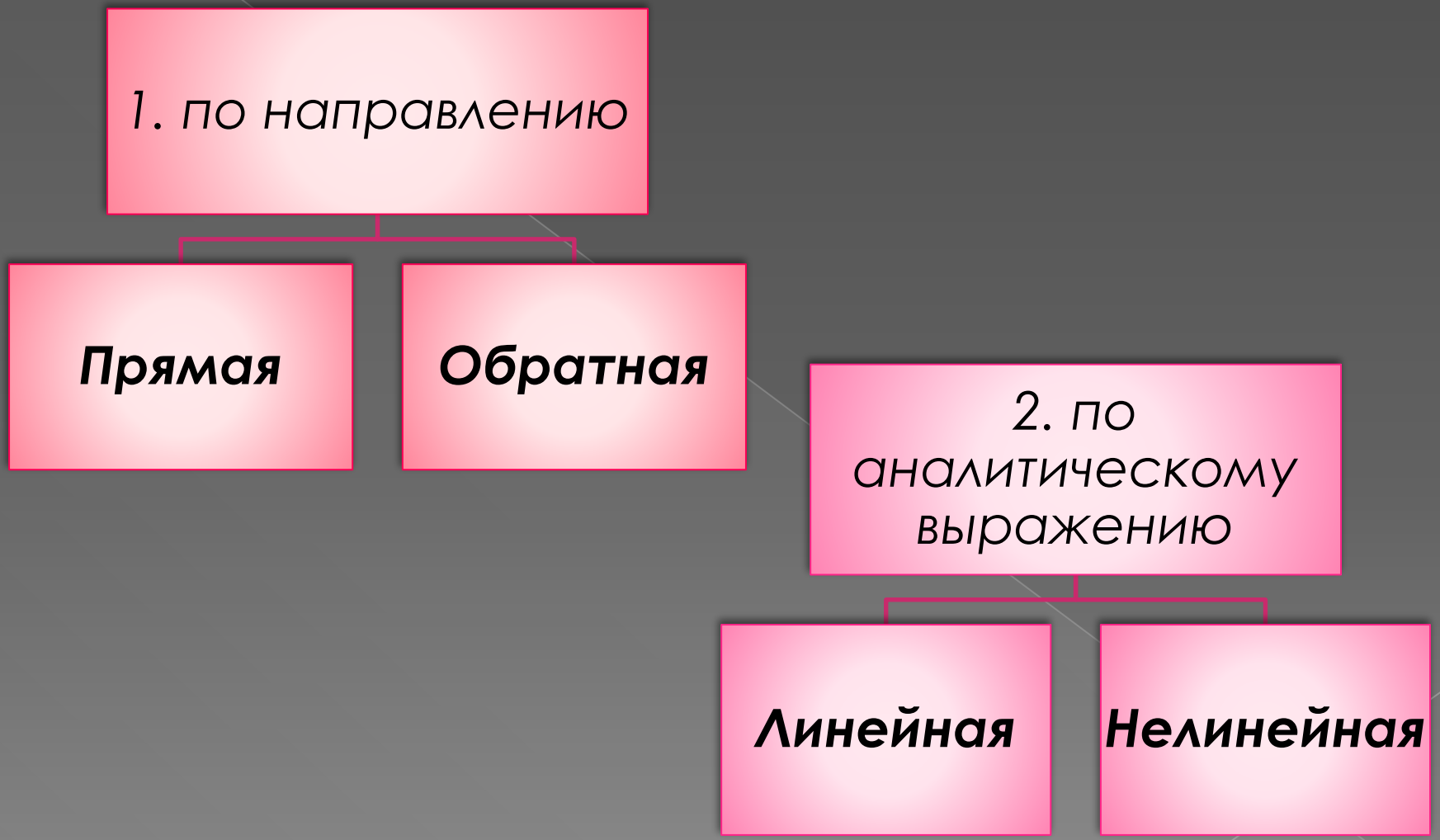
Корреляционно-регрессионный анализ

Включает в себя:

- 1. измерение **тесноты** и **направления** связи → корреляционный анализ

- 2. установление **аналитической формы** связи (формы зависимости признаков) → регрессионный анализ

Виды корреляционной связи



Корреляционный метод

- количественное определение тесноты связи между двумя признаками (при парной связи) и между результативным и множеством факторных признаков (при многофакторной связи).
- ***Корреляция*** - это статистическая зависимость между случайными величинами, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

Регрессионный метод

- аналитическое определение связи, в котором изменение одной величины (результативного признака) обусловлено влиянием одной или нескольких независимых величин (факторов), при этом множество прочих факторов, также влияющих на зависимую величину, принимается за постоянные и средние значения.

Регрессия

- По влиянию факторов
  - Однофакторная (парная)
  - Многофакторная (множественная)
- По форме зависимости
  - Линейная
  - Нелинейная
- По направлению связи
  - Прямая
  - Обратная

Парная корреляция и парная регрессия

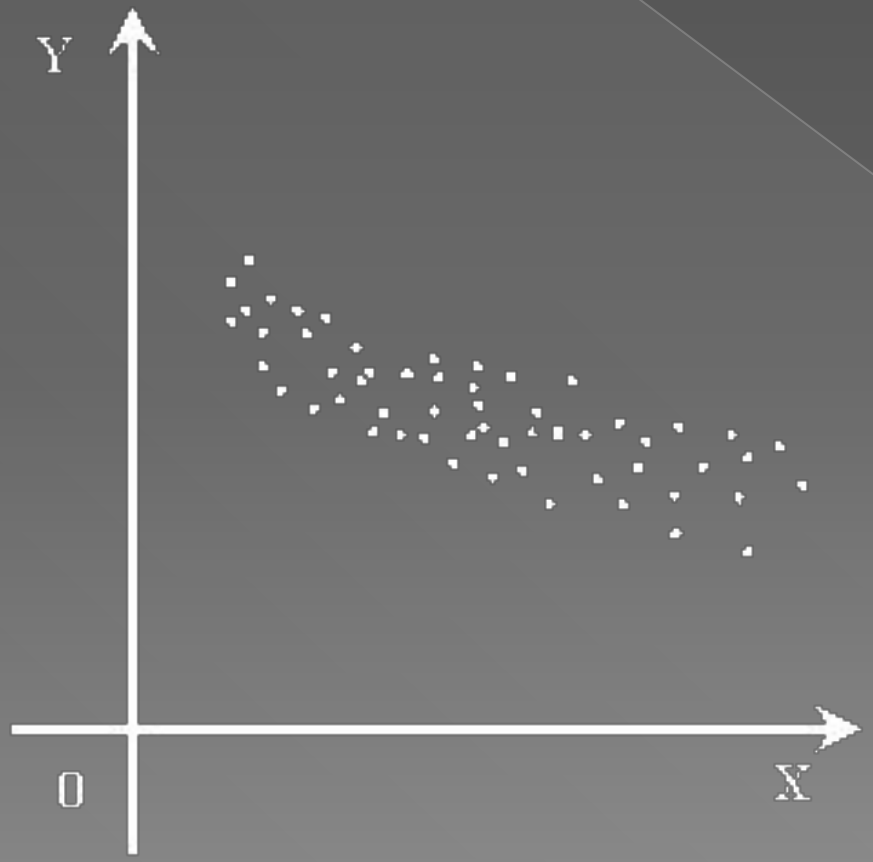
- характеризует связь между **двумя** признаками – результативным (Y) и факторным (X)

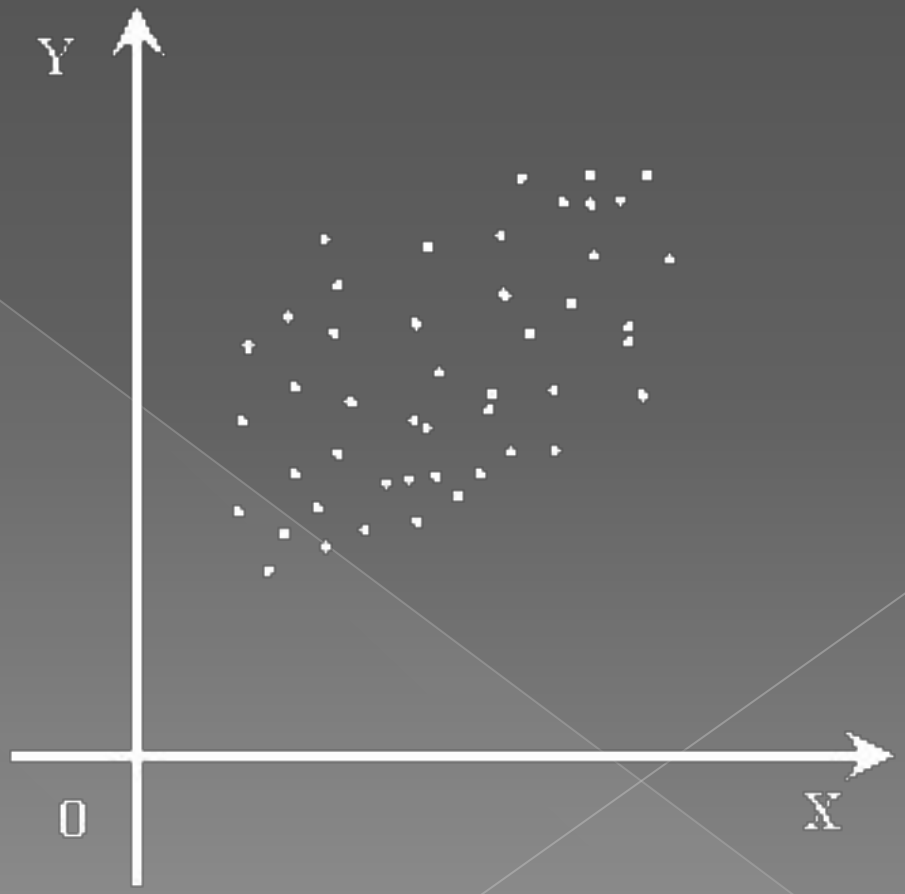
Простейший прием выявления связи между двумя признаками: **корреляционная таблица**

Приз-нак X/Y	Y_1	Y_2	…	Y_z	Итого	Y_i
X_1	f_{11}	…	…	f_{1z}	$\sum_{i}^{z} f_{1j}$	$\overline{Y_1}$
X_2	f_{21}	…	…	f_{2z}		$\overline{Y_2}$
…	…	…	…	…		…
X_r	…	…	…	…		…
Итого	$\sum_{i=1}^{k} f_{i1}$					$\overline{Y}$

- f – частота соответствующих сочетаний X и Y
- для каждого X_i рассчитывается среднее значение У

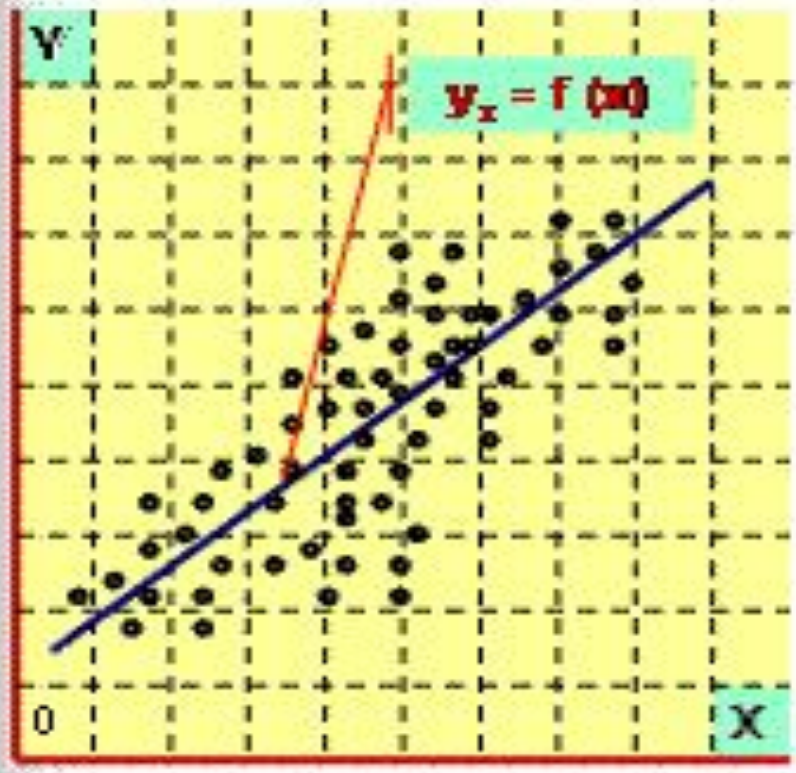
Множество точек $\{X_i, Y_i\}$ на плоскости XY - это **корреляционное поле**

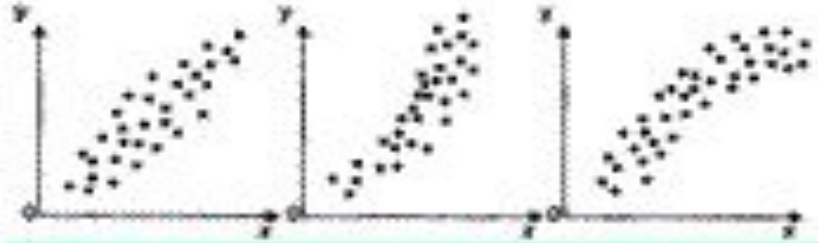




Корреляционное поле – аналитическая взаимосвязь

$$\bar{Y}_x = a_0 + a_1 x$$





Прямая (положительная) регрессия

Обратная (отрицательная) регрессия

$$\bar{Y}_x = a_0 + a_1 x + a_2 x^2$$

$$\bar{Y}_x = a_0 + \frac{a_1}{x}$$

Аналитическая связь (форма) между двумя признаками - описывается уравнениями:

$$\overline{Y}_x = a_0 + a_1 x$$

$$\overline{Y}_x = a_0 + a_1 x + a_2 x^2$$

Коэффициент корреляции

- Количественная оценка тесноты связи
- Характеризует тесноту и направление связи между двумя признаками в случае наличия между ними **линейной** зависимости

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}$$

Коэффициент корреляции принимает значения от -1 до +1

- $r > 0$ – связь прямая
- $r < 0$ – связь обратная
- $|r| < 0{,}30$ - связь слабая
- $|r| = 0{,}3 \ldots 0{,}7$ - связь средняя
- $|r| > 0{,}70$ - связь сильная, или тесная
- $|r| = 1$ - связь функциональная, т.е. каждому значению факторного признака строго соответствует одно значение результативного признака
- r близко к 0 - отсутствует линейная связь между У и Х (возможно нелинейное взаимодействие)

Метод наименьших квадратов

- Сущность метода: нахождение параметров модели (a_0, a_1), при которых минимизируется сумма квадратов отклонений фактических значений результативного признака от теоретических

$$\begin{cases} na_0 + a_1 \sum X = \sum y \\ a_0 \sum X + a_1 \sum X^2 = \sum Xy \end{cases}$$

Метод наименьших квадратов

- Сущность метода: нахождение параметров модели *(a, b)*, при которых минимизируется сумма квадратов отклонений **фактических** значений результативного признака от **теоретических**

$$\begin{cases} n \cdot a + b \cdot \sum X = \sum Y \\ a \cdot \sum X + b \cdot \sum X^2 = \sum (X \cdot Y) \end{cases}$$

Множественная (многофакторная) регрессия

- Изучение связи между тремя и более связанными между собой признаками.
- Аналитическое определение связи между результативным признаком и множеством факторных признаков, т.е нахождение функции:

$$\overline{Y} = f(X_1, X_2 \ldots X_n)$$

Метод перебора различных уравнений

- Сущность заключается в том, что большое число уравнений (моделей) регрессии, отобранных для описания связей какого-либо социально-экономического явления или процесса, реализуется на ЭВМ с помощью специально разработанного алгоритма перебора с последующей статистической проверкой, главным образом, на основе t - критерия Стьюдента и F-критерия Фишера-Снедекора.

Оценка значимости и правильности параметров взаимосвязи

- Получив значения корреляции и уравнение регрессии - необходимо проверить их на соответствие **истинным параметрам взаимосвязи:**
- При этом:
- **1) значимость коэффициента корреляции проверяется** на основе
  t– критерий Стьюдента
- **2) правильность выбора вида взаимосвязи и всего уравнения регрессии** проверяется на основе
  F– критерий Фишера

Оценка значимости **параметров** корреляции: *t– критерий Стьюдента*

- **Значимость коэффициента корреляции** проверяется на основе t- критерия Стьюдента.
- При этом выдвигается и проверяется нулевая гипотеза о равенстве коэффициента корреляции нулю (r=0), т.е гипотеза об отсутствии взаимосвязи.
- При проверке этой гипотезы используется t-статистика (из специальных таблиц).

Оценка значимости параметров взаимосвязи: *t– критерий Стьюдента*

$$t_{расч} = r_{xy} \cdot \sqrt{\frac{n-2}{1-r_{xy}^2}},$$

- где $t_{расч}$ –расчетное значение t-критерия.
- Если расчетное значение $t_р > t_{кр}$ (табличное), то гипотеза отвергается , что свидетельствует о значимости линейного коэффициента корреляции, а следовательно, и о статистической зависимости между X и У

Оценка значимости параметров взаимосвязи: *F– критерий Фишера*

- Вывод о правильности выбора **вида взаимосвязи** и **характеристику значимости всего уравнения** регрессии получают с помощью F-критерия.
- При этом выдвигается и проверяется нулевая гипотеза о несоответствии заложенных в уравнении регрессии связей реально существующим.

Оценка правильности выбора вида взаимосвязи и уравнения регрессии : *F– критерий Фишера*

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

- Если Fp>Fa при a = 0,05 или a = 0,01, то гипотеза о несоответствии заложенных в уравнении регрессии связей реально существующим отвергается.
- Величина Fa определяется по специальным таблицам, входом в которые являются величины a = 0,05 или a = 0,01 и числа степеней свободы: v1 = k -1, v2 = n -k, где n - число наблюдений, k - число факторных признаков в уравнении

Измерение связи между **качественными** признаками

Таблица взаимной сопряженности

C - число объектов, обладающих свойством A,
но не обладающих свойством B в изучаемой совокупности

	A	A	Сумма
B	a	b	a+b
в	c	d	c+d
Сумма	a+c	b+d	

Коэффициент ассоциации

- Зависимости
- a, b, c, d - значения признаков в клетках матрицы сопряженности альтернативных признаков

$$K_{ac} = \frac{ad - bc}{ad + bc}.$$

Коэффициент ассоциации принимает значения от -1 до +1

- Если коэффициент имеет положительный знак (+), то связь положительная,
- Если коэффициент имеет отрицательный знак (-), то связь отрицательная.
- $K_{ac} = 0$ - корреляция отсутствует (данные факторы между собой нейтральны);
- $0,09 \leq K_{ac} \leq 0,19$ - статистическая взаимосвязь очень слабая;
- - если $0,2 \leq K_{ac} \leq 0,49$ - статистическая взаимосвязь слабая;
- - если $0,5 \leq K_{ac} \leq 0,69$ - статистическая взаимосвязь средняя;
- - если $0,70 \leq K_{ac} \leq 0,99$ - статистическая взаимосвязь сильная.

Коэффициент контингенции (сопряженности)

- В отличие от коэффициента ассоциации он учитывает двустороннюю взаимосвязь альтернативных признаков, т.е. производит измерение более чутко и корректно.

$$K_{\text{кон}} = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}.$$

Коэффициент детерминации

- зависимости

Эластичность

- зависимости