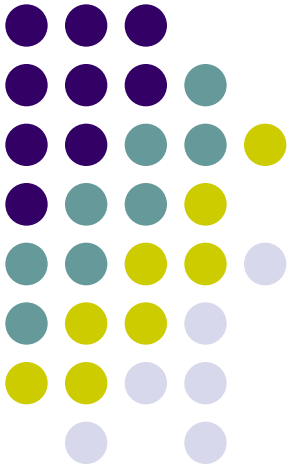
КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

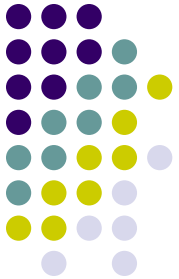
АНАЛИЗ ПАРНЫХ ВЗАИМОСВЯЗЕЙ



Основные понятия

- ***Статистическая*** связь и ее отличие от ***функциональной.***
- Связь как ***синхронность*** (согласованность) – корреляционный анализ.
- Связь как ***зависимость*** (влияние) – регрессионный анализ (причинно-следственные связи).
- ***Парная связь*** как частный случай множественной связи.
- Неучтенные факторы.

Этапы анализа



- Выявление наличия взаимосвязи между признаками
- Определение формы связи
- Определение силы (тесноты) и направления связи

Выявление наличия связи между признаками

Диаграммы рассеяния

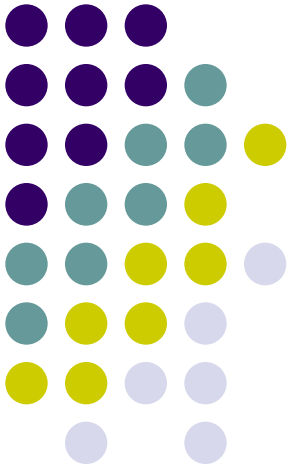
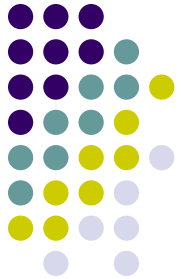
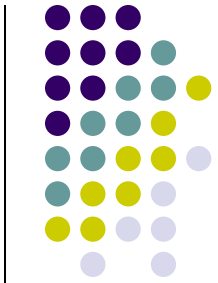


Диаграмма рассеяния (scatterplot)



Направление связи



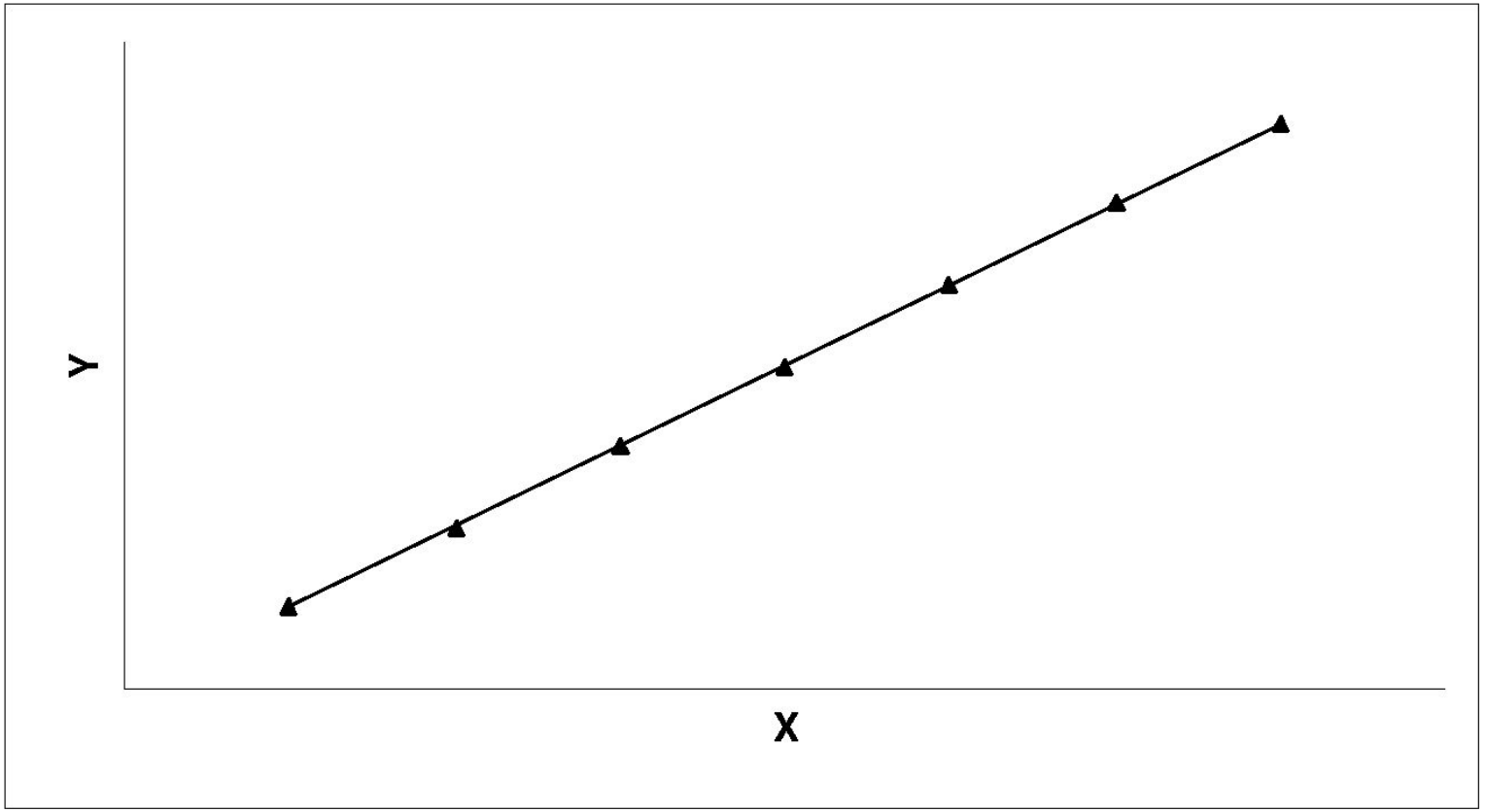
- В случае положительной функциональной связи –

  чем больше значения одного признака, ***тем больше*** значения другого и

  чем меньше значения одного признака, ***тем меньше*** значения другого.

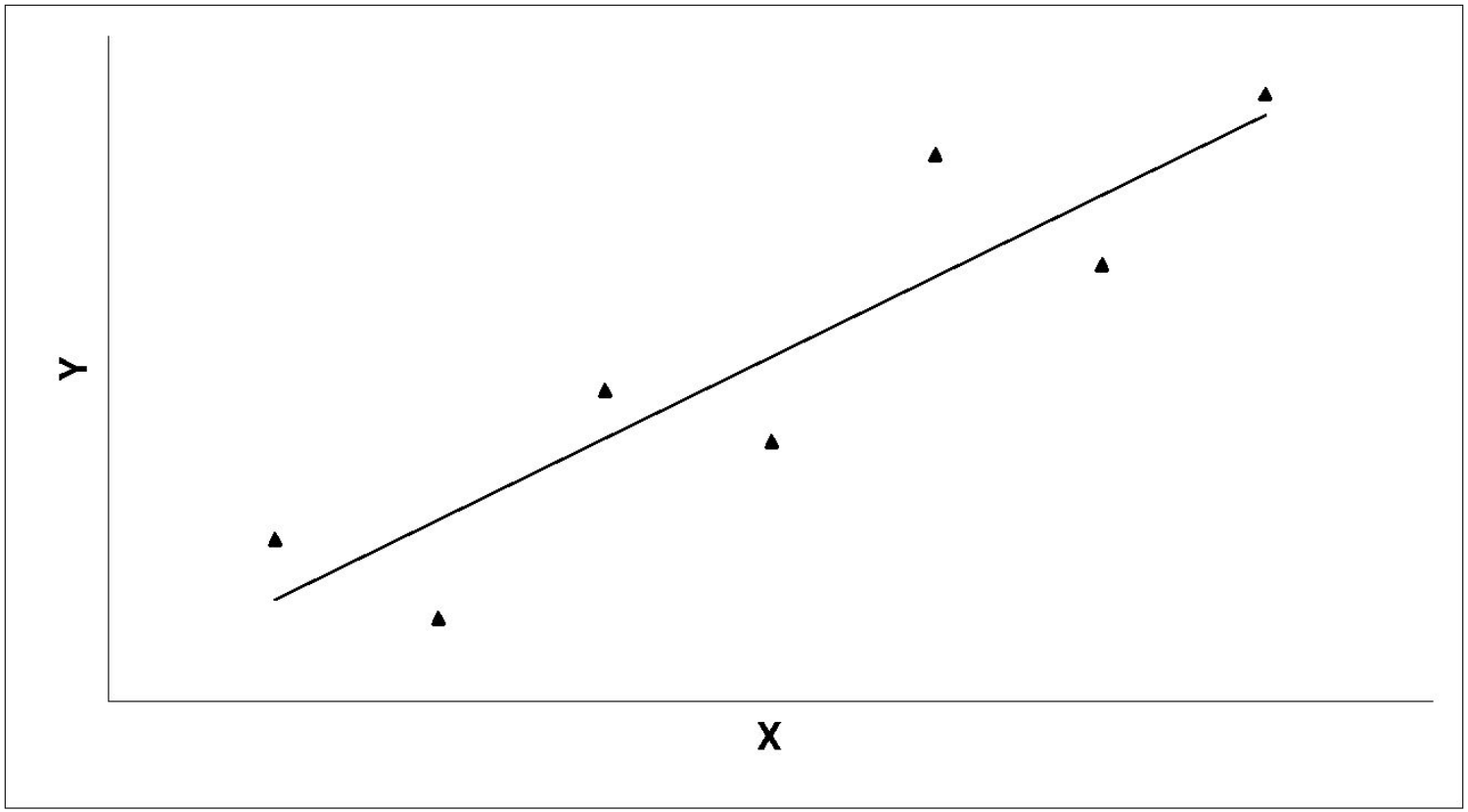
Направление связи

- Пример положительной функциональной связи между признаками X и Y.



Направление связи

- Пример положительной статистической связи между признаками X и Y.



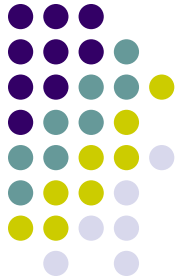
Направление связи

- В случае положительной статистической связи –

  чем больше значения одного признака,
  тем больше <u>в среднем</u> значения другого и

  чем меньше значения одного признака,
  тем меньше <u>в среднем</u> значения другого.

Направление связи



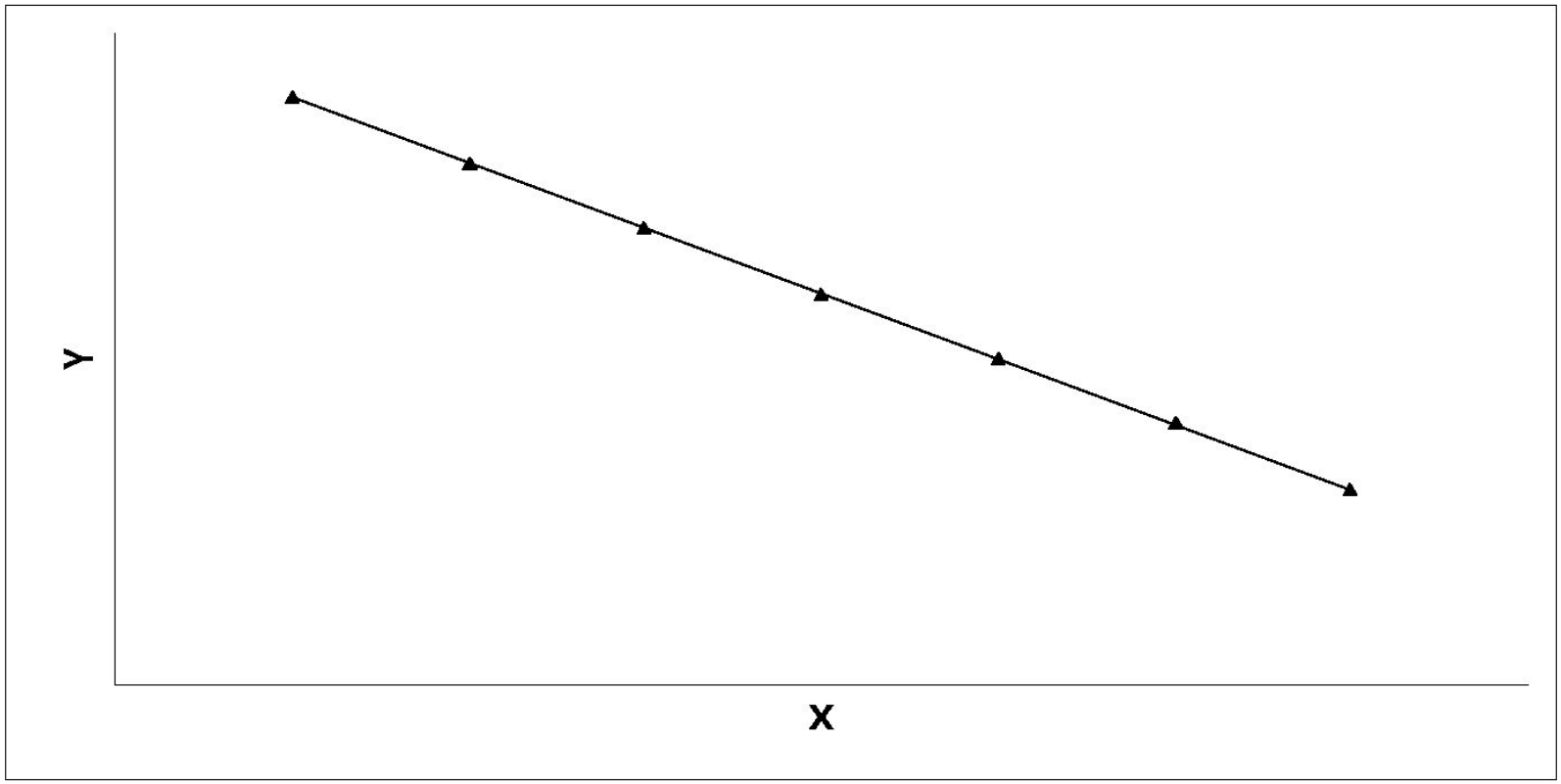
- В случае отрицательной функциональной связи –

  чем больше значения одного признака, ***тем меньше*** значения другого и

  чем меньше значения одного признака, ***тем больше*** значения другого.

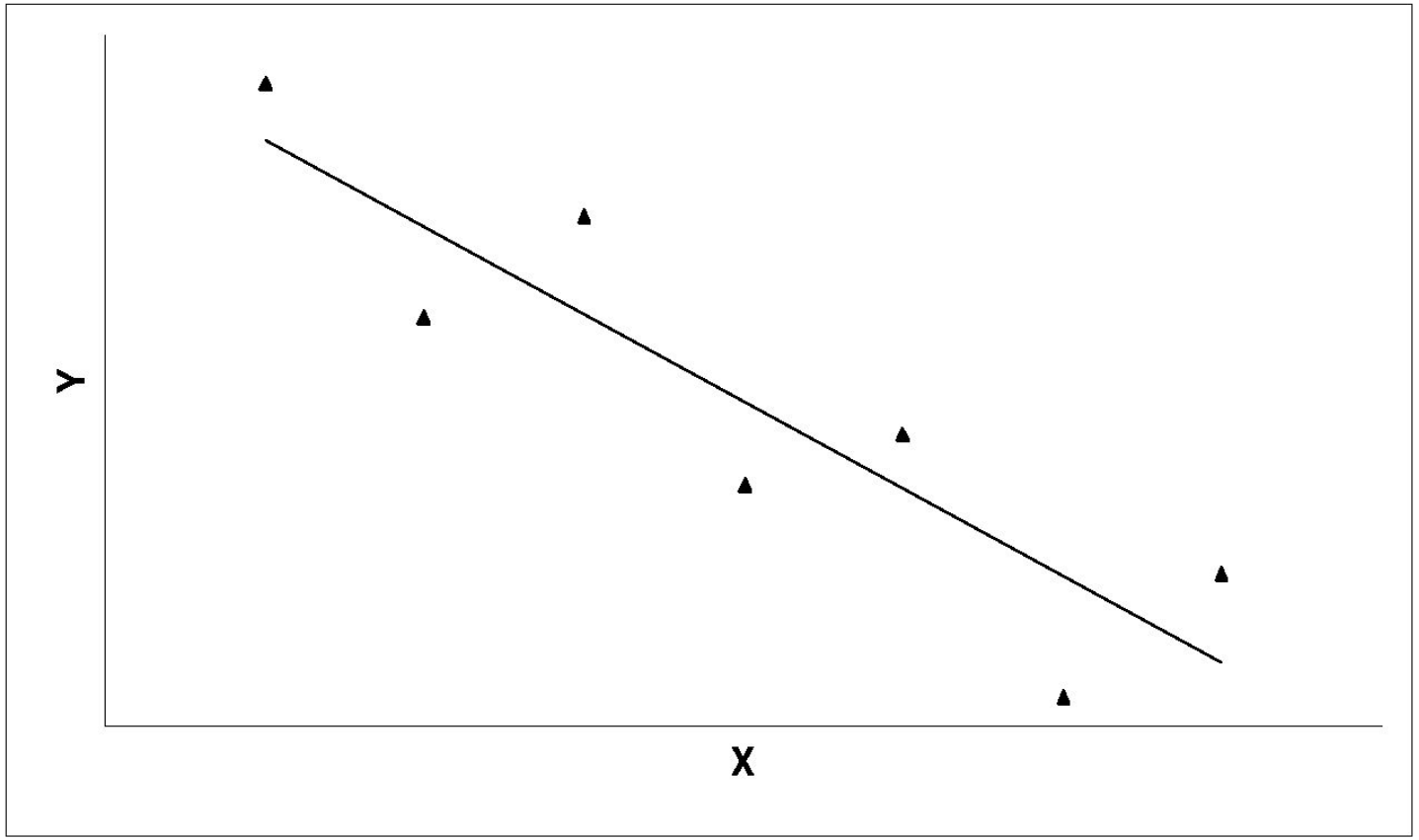
Направление связи

- Пример отрицательной функциональной связи между признаками X и Y.



Направление связи

- Пример отрицательной статистической связи между X и Y.



Направление связи

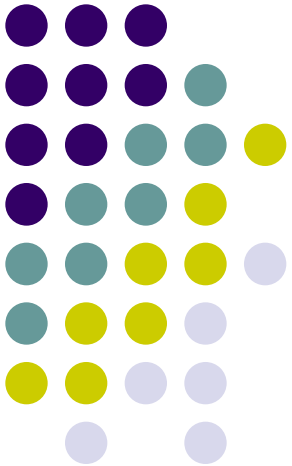
- В случае отрицательной статистической связи –

  чем больше значения одного признака,
  тем меньше <u>в среднем</u> значения другого и

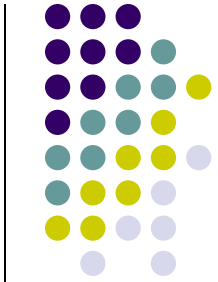
  чем меньше значения одного признака,
  тем больше <u>в среднем</u> значения другого.

Подбор формы связи

Линейная связь



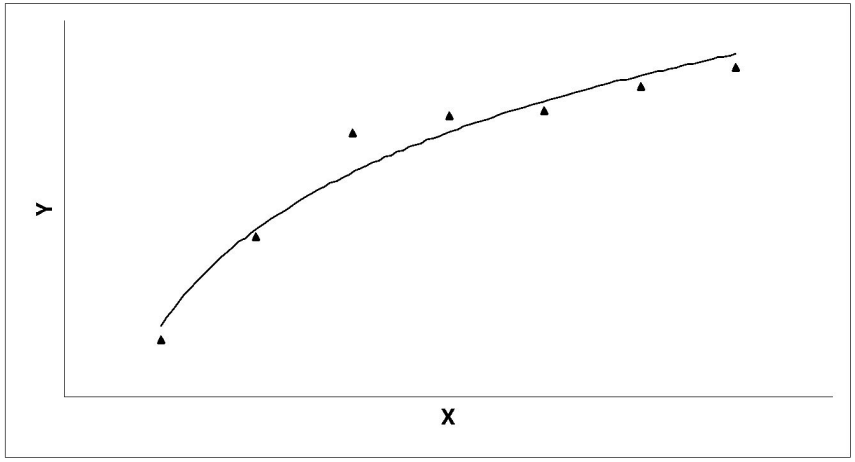
Форма связи



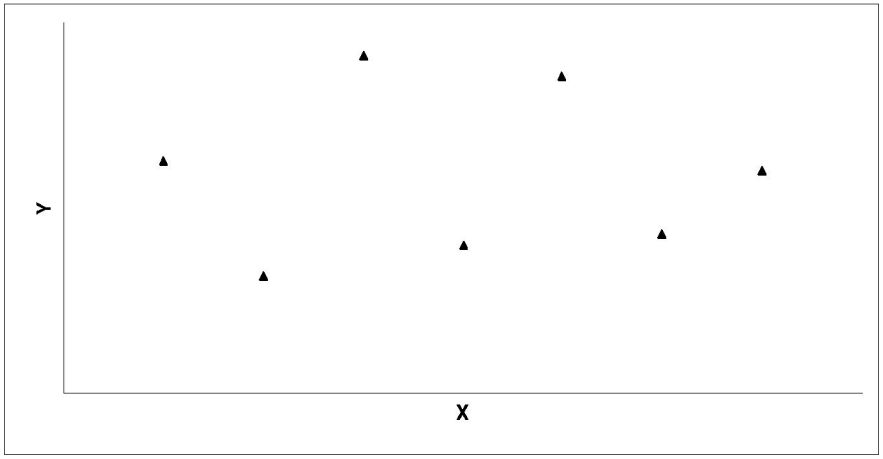
- ***Почему прямая?***
- Поскольку наиболее простой формой зависимости в математике является прямая, то в корреляционном и регрессионном анализе наиболее популярны ***линейные модели***.

Форма связи

- Примеры нелинейной связи (рис. а) и отсутствия связи (рис. б) между признаками X и Y

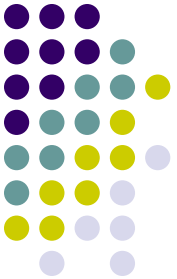


а

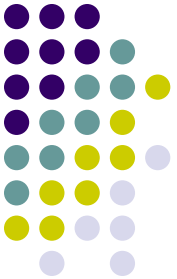


б

Форма связи – ?

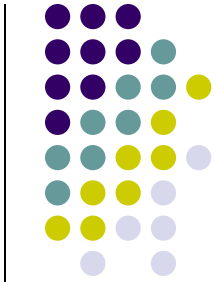


Форма связи



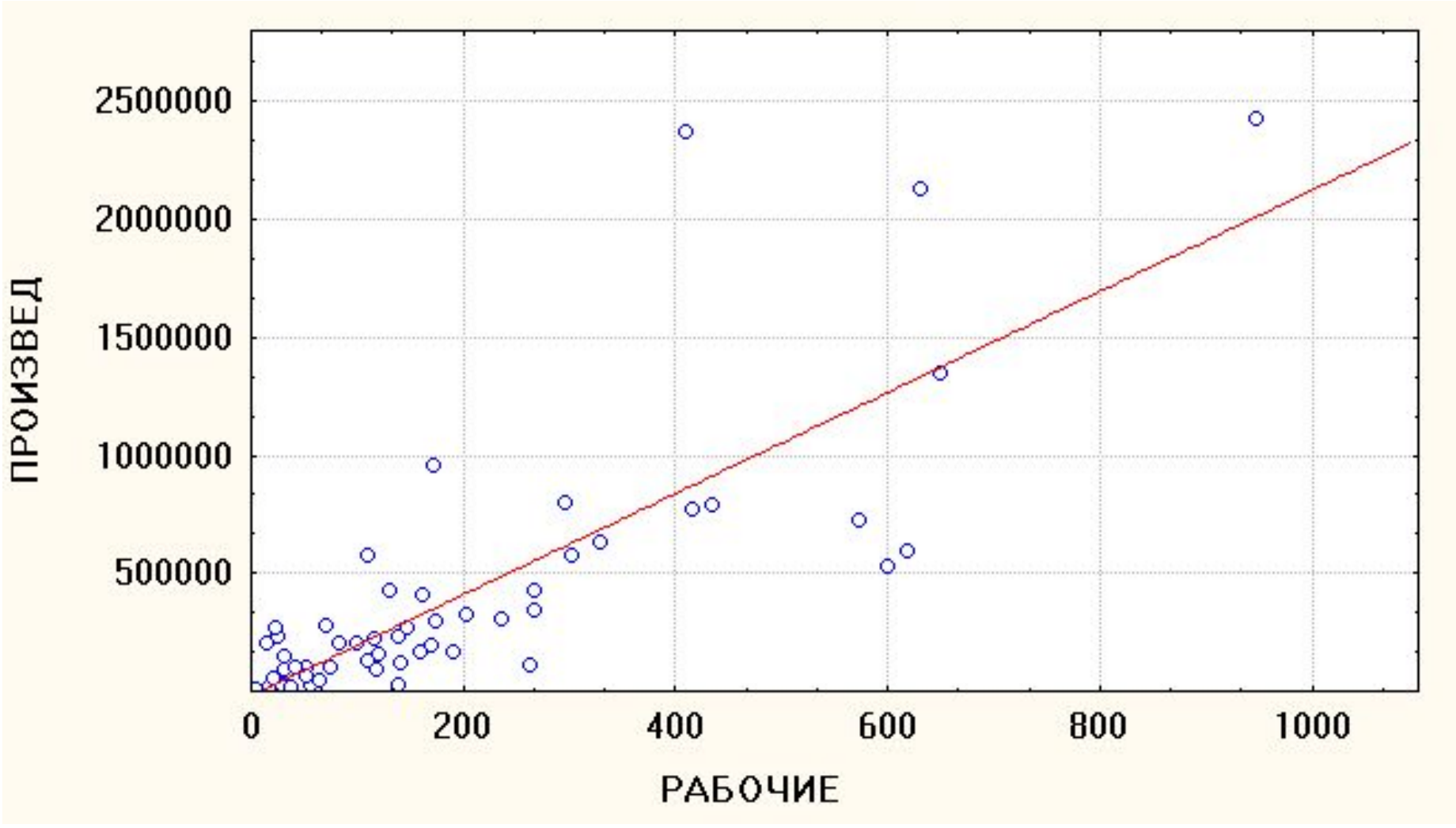
- Сколько прямых можно провести через облако точек на диаграмме рассеяния?
- Есть ли среди них наилучшая?
- Каким методом ее можно найти?

Форма связи



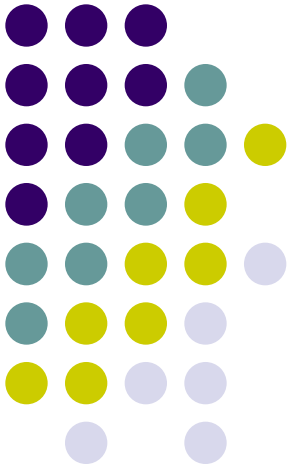
- ***Метод наименьших квадратов*** позволяет построить наилучшую прямую – ***линию регрессии***.
- Сумма квадратов расстояний от точек до этой линии минимальна (по сравнению со всеми возможными линиями).

Линия регрессии

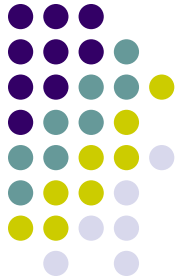


Коэффициент корреляции

Мера тесноты линейной связи



Коэффициент корреляции

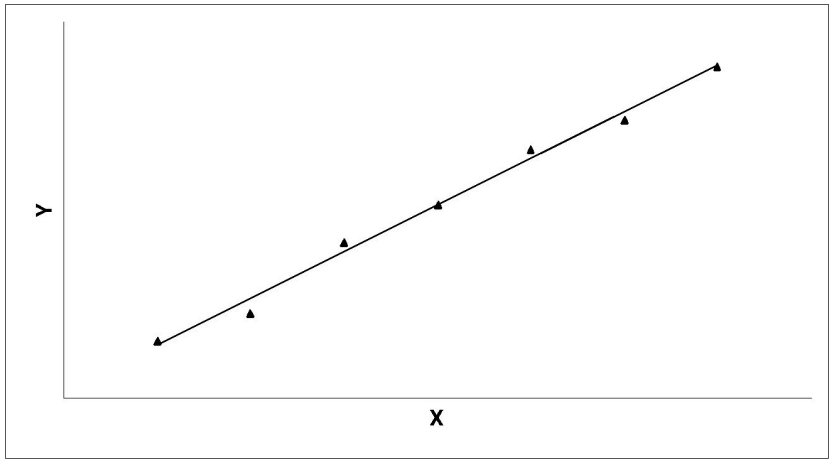


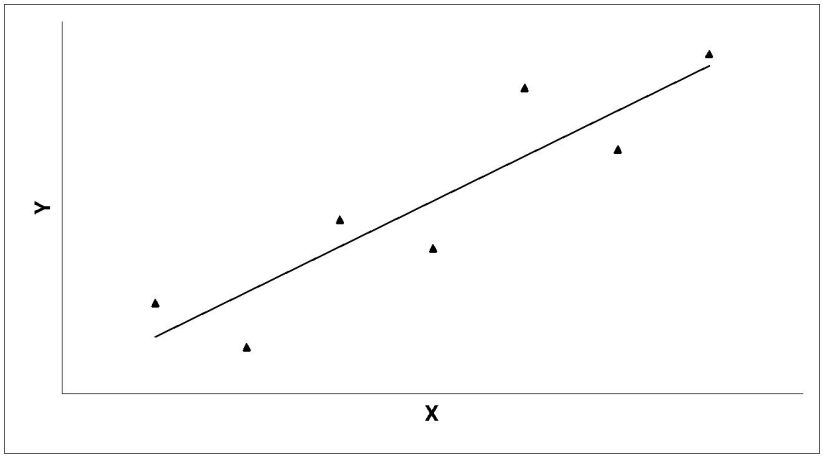
- Оказывается форма связи (линия регрессии) не дает ответа на вопрос о тесноте (силе) связи пары переменных.

- На вопрос о силе связи отвечает ***коэффициент парной корреляции***. Он показывает, насколько тесно две переменные связаны между собой.

Коэффициент корреляции

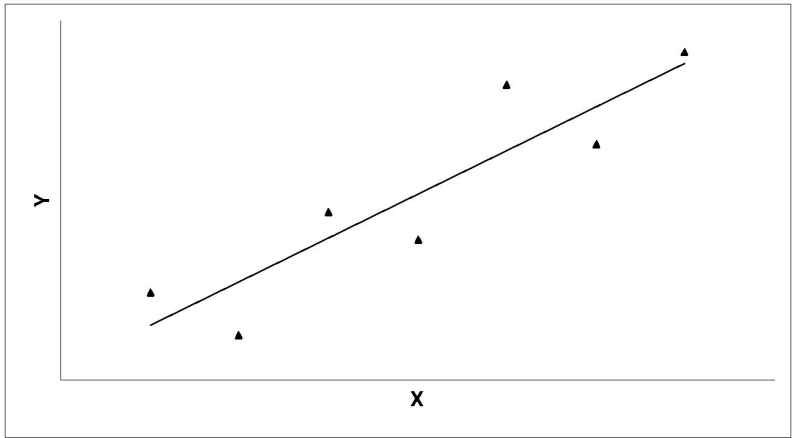
- На каком из двух графиков связь между признаками сильнее (теснее), т.е. какому из графиков соответствует более высокий коэффициент корреляции?

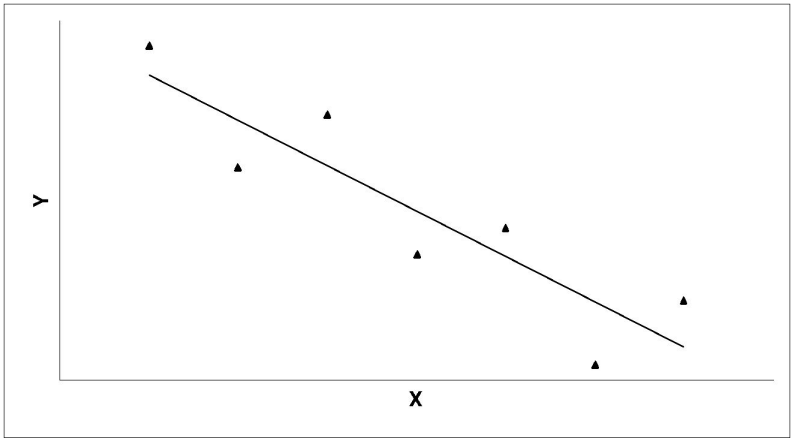




Коэффициент корреляции

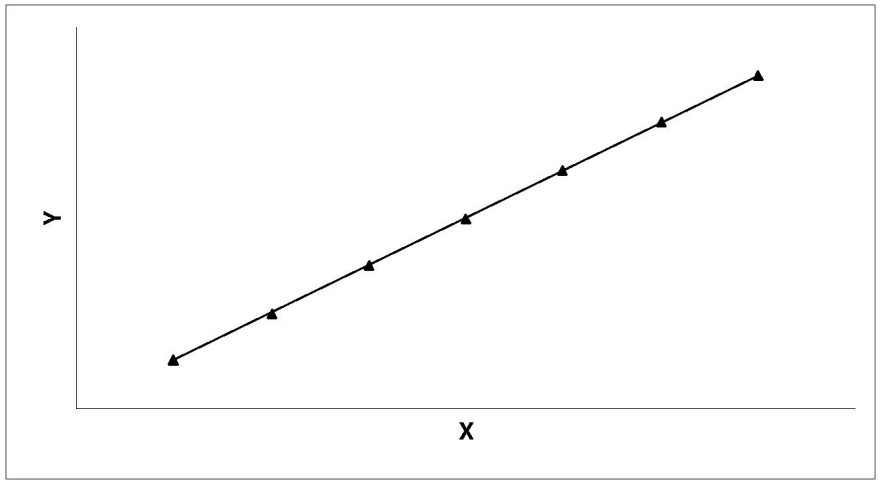
- Коэффициент парной корреляции ***r*** принимает значения в диапазоне от –1 до +1.
- Положительные значения коэффициента корреляции ***r*** свидетельствуют о положительной связи между признаками, отрицательные – об отрицательной связи.





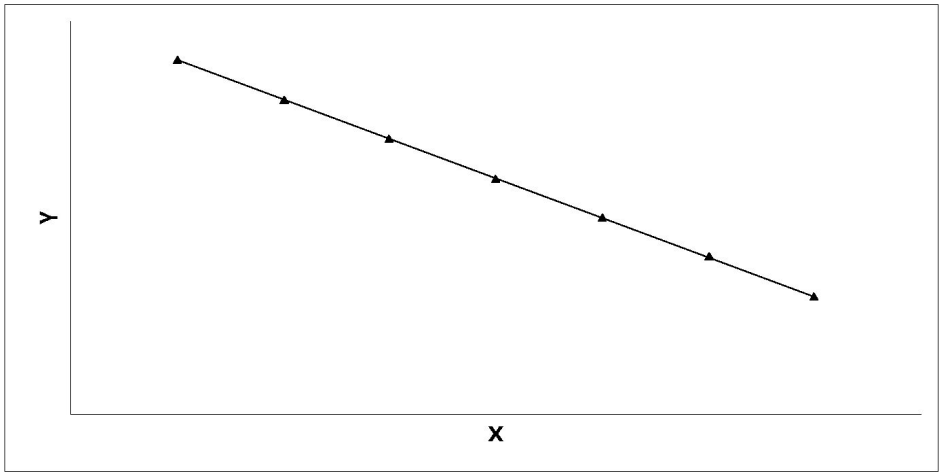
Коэффициент корреляции

- Между двумя переменными существует ***функциональная положительная линейная связь***.
- ***r* = ?**



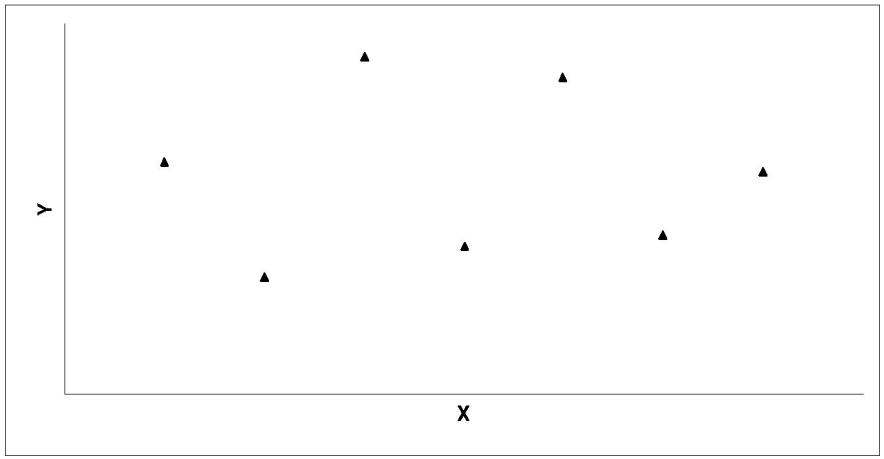
Коэффициент корреляции

- Между двумя переменными существует ***функциональная отрицательная линейная связь***.
- ***r* = ?**

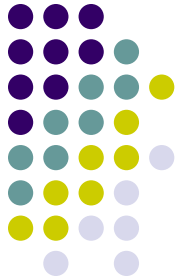


Коэффициент корреляции

- Переменные **линейно независимы**, т.е. на диаграмме рассеяния облако точек "вытянуто по горизонтали".
- ***r* = ?**



Коэффициент корреляции



- Визуально о силе связи можно судить по тому, насколько тесно расположены точки-объекты около линии регрессии.

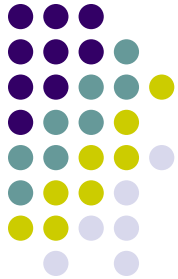
  Чем ближе точки к линии регрессии, тем сильнее связь.

Коэффициент корреляции

- Формула для вычисления парного коэффициента корреляции:

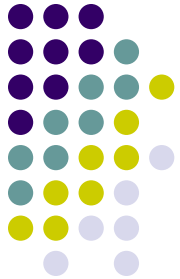
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Коэффициент корреляции



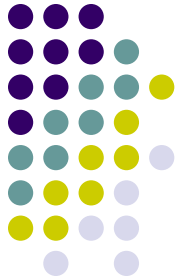
- Коэффициент парной корреляции вычисляется для количественных признаков.
- Коэффициент корреляции симметричен, т.е. не изменяется, если X и Y поменять местами.
- Коэффициент корреляции является величиной ***безразмерной***.
- Коэффициент корреляции ***не изменяется при изменении единиц измерения*** признаков X и Y.

Коэффициент детерминации



- Для интерпретации результатов корреляционного анализа обычно используется ***коэффициент детерминации d*** (***d*** = r^2, выражается в %)
- Коэффициент детерминации показывает, насколько изменения зависимого признака объясняются изменениями независимого

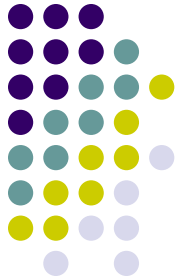
Коэффициенты корреляции и детерминации



- Коэффициент детерминации принимает значения в диапазоне от 0% до 100%.

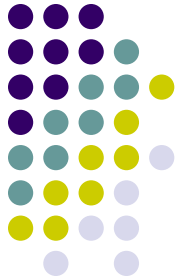
- Если две переменные функционально линейно зависимы, что можно сказать о коэффициенте детерминации?
- Чему при этом равен коэффициент корреляции?

Коэффициенты корреляции и детерминации



- Если две переменные линейно независимы, что можно сказать о коэффициенте детерминации?
- А о коэффициенте корреляции?

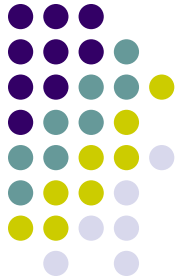
Коэффициенты корреляции и детерминации



- Чем выше по модулю (по абсолютной величине) значение коэффициента корреляции, тем сильнее связь между признаками.

- Если $|r| > 0.7$, связь называется сильной;
  если $0,5 < |r| \leq 0,7$ – средней;
  если $|r| \leq 0,5$ – слабой.

Матрица корреляции



- Если объекты характеризуются несколькими признаками, можно построить ***матрицу корреляции***.

- По диагонали матрицы стоят ???
- Матрица симметрична, т.е. значения выше и ниже диагонали повторяются (т.к. $r_{xy} = r_{yx}$). Почему?

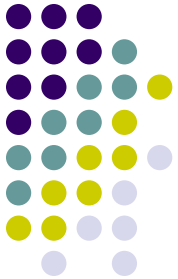
Матрица корреляции

- Пример матрицы корреляции для трех признаков.

Correlations (INDUSTRY.STA)
Marked correlations are significant at p < ,05000
N=1060 (Casewise deletion of missing data)

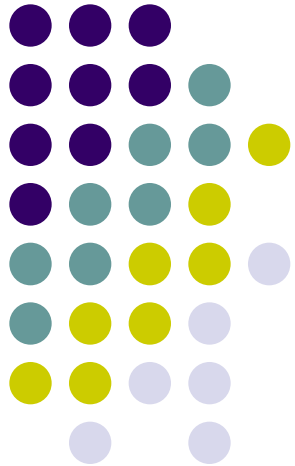
Variable	**РАБОЧИЕ**	ПРОИЗВЕД	ДВИГАТЕЛ
РАБОЧИЕ	1,00	0,52	0,43
ПРОИЗВЕД	0,52	1,00	0,30
ДВИГАТЕЛ	0,43	0,30	1,00

Матрица корреляции

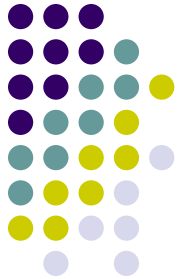


- Некоторые коэффициенты в матрице корреляции показаны красным цветом.
- Это означает, что они являются ***статистически значимыми***.

Значимость коэффициента корреляции

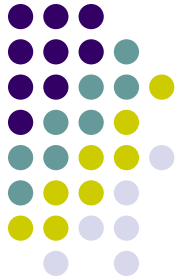


Статистическая значимость коэффициента корреляции



- Если коэффициент корреляции вычислен на основе выборки, то возможны две гипотезы:
  - он отражает связь, которая действительно существует в генеральной совокупности;
  - он объясняется случайным эффектом выборки, а в генеральной совокупности коэффициент корреляции равен нулю, т.е. (линейной) связи нет.
- Какая гипотеза верна?

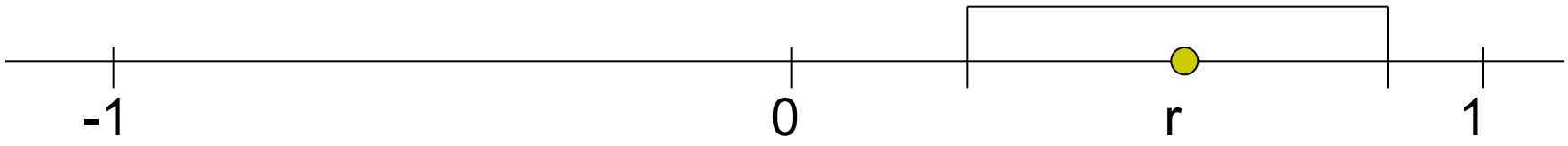
Статистическая значимость коэффициента корреляции



- Надо понять, ***как далеко значение r от нуля***.
  - Для построения доверительного интервала вычисляется стандартная ошибка ***r***.
  - Затем она умножается на параметр *t*, зависящий от доверительной вероятности *P,* чтобы найти предельную ошибку.
  - Наконец, строится доверительный интервал для возможных значений ***r*** в генеральной совокупности.
- Остается проверить, попадет ли нулевое значение в этот интервал.

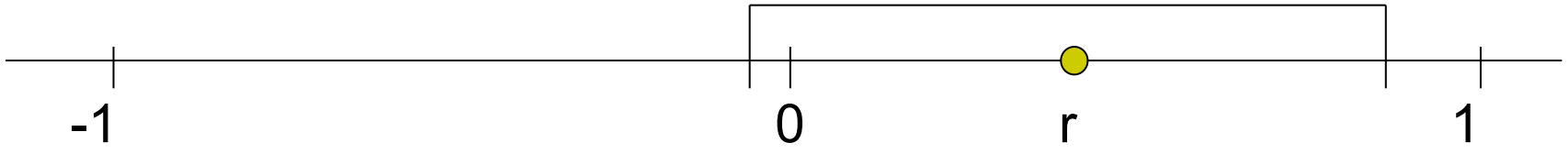
Статистическая значимость коэффициента корреляции

- Если ***ноль не попадет в доверительный интервал***, значит с высокой вероятностью в генеральной совокупности не может быть нулевого значения коэффициента корреляции, т.е. связь между признаками существует и в генеральной совокупности. В таком случае ***коэффициент корреляции является статистически значимым***.

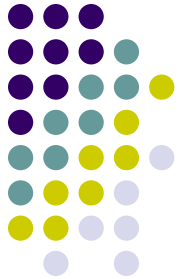


Статистическая значимость коэффициента корреляции

- Если ***ноль попадет в доверительный интервал***, значит с высокой вероятностью в генеральной совокупности может оказаться нулевая корреляция, т.е. отсутствие связи. В таком случае ***коэффициент корреляции является статистически незначимым***.



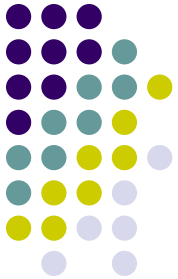
Статистическая значимость коэффициента корреляции



- На практике незначимые коэффициенты можно считать нулями и принимать во внимание только значимые.

- Величина коэффициента корреляции еще не гарантирует его значимости.

Статистическая значимость коэффициента корреляции



- Может ли большой коэффициент корреляции оказаться статистически незначимым?
  При каких условиях?

- Может ли небольшой коэффициент корреляции оказаться статистически значимым?
  При каких условиях?