Лекция по дисциплине

«Малозатратные инженерные технологии»

автор: к.т.н., доц. Тимошек Игорь Николаевич

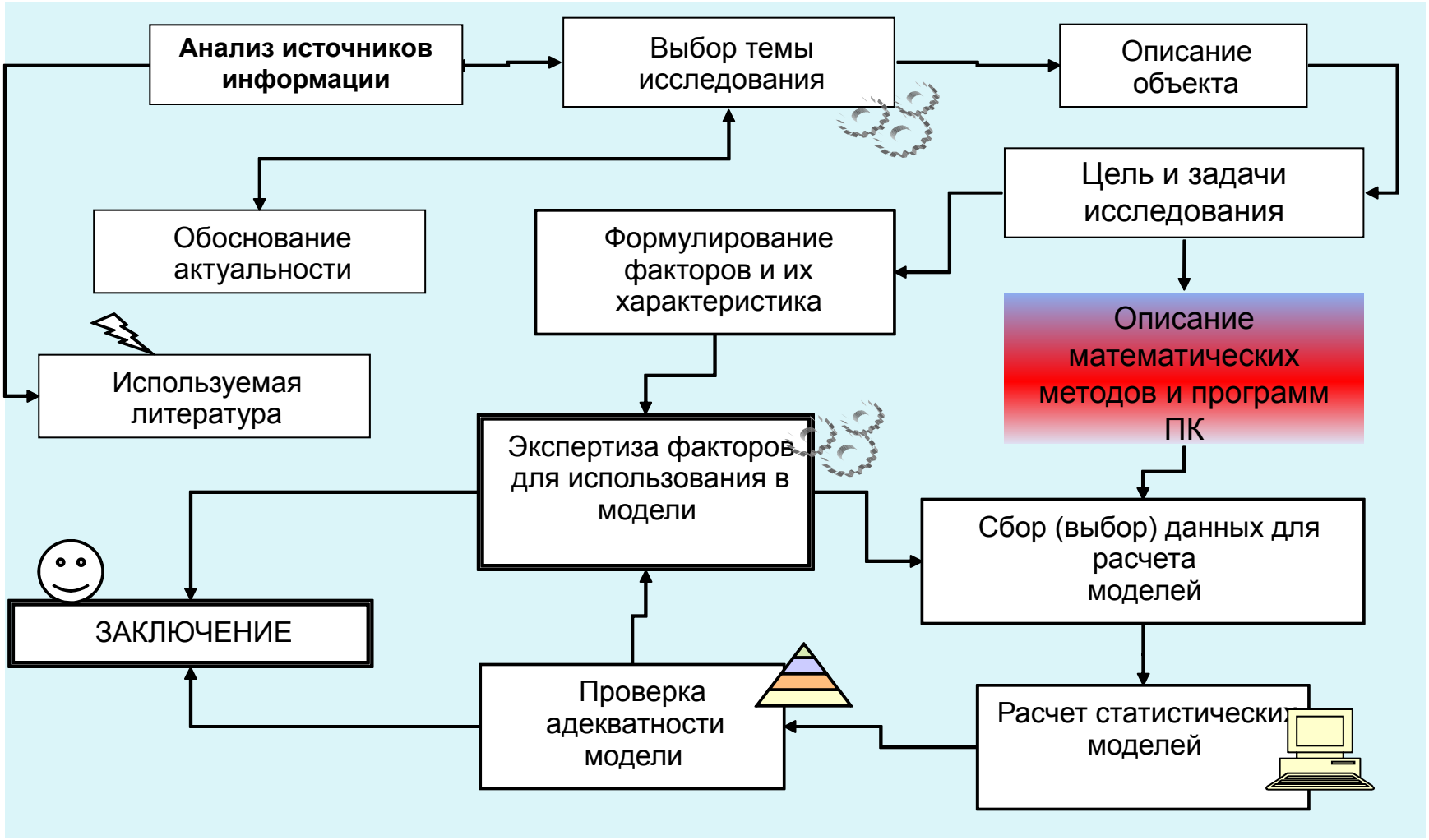
Тема: «Основные математические положения, применяемые для анализа и построения статистической модели»

Вопросы: ***1. Основы корреляционного анализа***

2. Основы регрессионного анализа

3. Методы, принципы и алгоритмы формирования модели

Основные этапы выполнения расчетной работы



Цель занятия:

- **Припоминание** основ статистического моделирования (изученного на втором курсе в математике и статистике) процессов для анализа функционирования логистических объектов и на их основе выработки практических рекомендаций по совершенствованию исследуемых объектов.

«Тем, кто понимает суть регрессии и корреляции, советы не нужны. Тем, кто не понимает, никакие советы не помогут.»

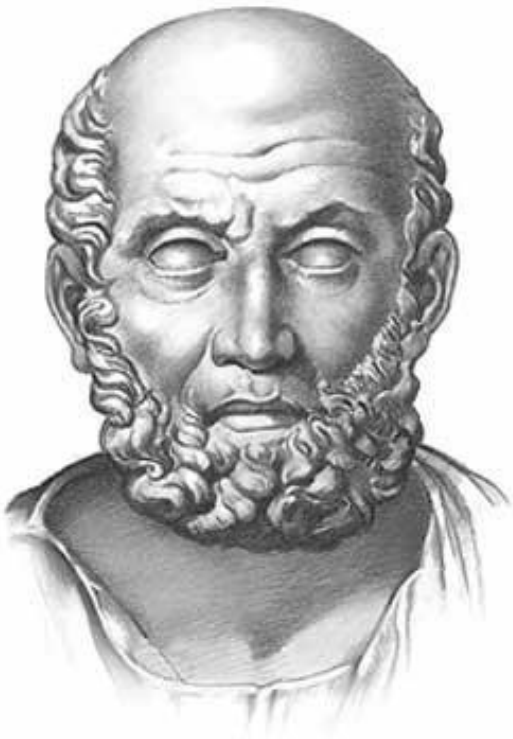
(Из книги Н. Джонсона и Ф. Лиона «Статистика и планирование эксперимента в технике и науке»).

Описание проблемы:

перед исследователями всегда стояла задача достоверного отображения объективно существующих закономерностей в деятельности транспортного предприятия для конкретных условий, в которых оно реализует свою деятельность (перевозку, складирование, хранение грузов и товаров) с обязательной количественной оценкой причинно-следственных взаимосвязей многообразия факторов.

Корреляционный анализ. История

- Еще Гиппократ (греческий врач и педагог, чье имя связывается в представлении большинства людей со знаменитой клятвой, которая символизирует высокие этические нормы европейской медицины)обратил внимание на то, что между **телосложением и темпераментом людей**, между строением их тела и предрасположенностью к заболеваниям **существует определенная связь**.



ГИППОКРАТ
460-370 до н. э.

Корреляционный анализ. История

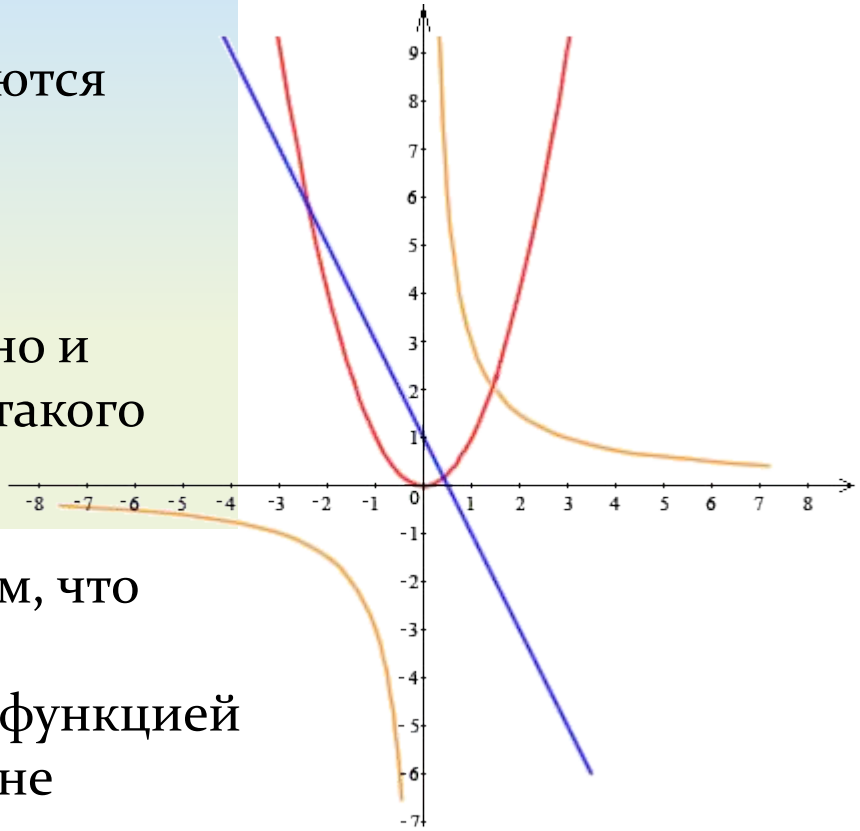
- Так современные **логистические исследования**, транспортных процессов посвящены установлению **закономерностей** между достигаемым **результатом** и целым рядом технических, технологических, эксплуатационных, психофизических, метеорологических и множество других **характеристик** (скорость перемещения объекта, производительность погрузки или разгрузки и отдельных звеньев объекта, показатели эксплуатационной надежности, объемы хранения и энергозатраты, связанные с функционированием объекта, и др.).
- При этом постоянно возникает вопрос о **взаимосвязи отдельных признаков**.

Корреляционный анализ. История

- Этой цели служит математическое понятие функции, имеющее в виду случаи, когда определенному значению одной (независимой) переменной X, называемой ***аргументом***, соответствует определенное значение другой (зависимой) переменной Y, называемой ***функцией***. Однозначная зависимость между переменными величинами *Y* и *X* называется ***функциональной***, т.е. $Y = f(X)$.
- Например, в функции $y = -2x+1$ каждому значению *x* соответствует в определенное значение *y*.
- В функции $y = x^2$ каждому значению *Y* соответствует 2 определенных значения *X*. *Графически это выглядит так (см. рисунок):*

Примеры элементарных функций:

- Но такого рода **однозначные** или **функциональные** связи между переменными величинами встречаются кране **редко**.
- Известно, что **новый** погрузочный механизм должен иметь большую **производительность**, чем морально и физически устаревший. Однако из такого правила бывают исключения.



- Причина таких «исключений» в том, что каждый признак, выражаясь математическим языком, является функцией многих переменных; на его величине сказывается влияние и других факторов, в том числе и случайных, что вызывает варьирование признаков.

Корреляционный анализ. История

- Отсюда зависимость между величинами приобретает не функциональный, а ***статистический характер***, когда определенному значению одного признака, рассматриваемого в качестве независимой переменной, соответствует не одно и то же числовое значение, а целая гамма распределяемых в вариационный ряд числовых значений другого признака, который рассматриваемого в качестве независимой переменной. Такого рода зависимость между переменными величинами называется ***корреляционной*** или ***корреляцией.***
- Термин «корреляция» происходит от лат. correlatio — соотношение, связь). При этом *данный вид взаимосвязи между признаками проявляется в том, что при изменении одной из величин изменяется среднее значение другой*.

Корреляционный анализ. История

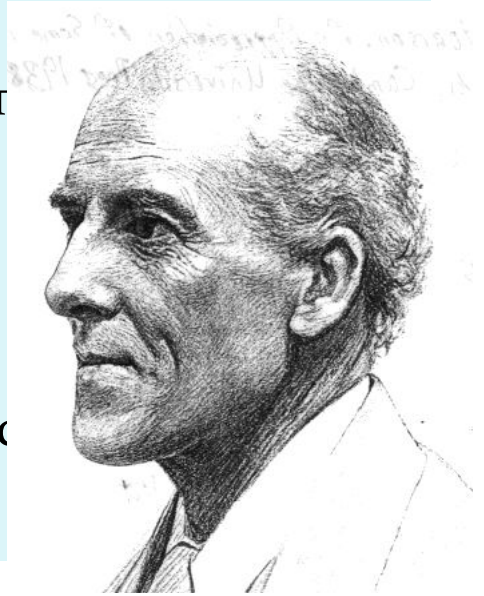
- *Слово «**статистика**» приходит от латинского слова status (состояние), которое употреблялось в значении «политическое состояние». Отсюда итальянские слова stato – государство и statista – знаток государств, отсюда также и немецкое слово Staat и английское state. В научный оборот слово «статистика» ввёл профессор Геттингенского университета **Готфрид Ахенваль** (1719 - 1772), понималось оно тогда как государствоведение.*
- В первой половине 19 века возникло **статистико-математическое направление** данной науки. Среди представителей этого направления следует отметить бельгийского статистика **Адольф Кетле** (1796-1874 гг.) – основоположника учения о средних величинах.

Корреляционный анализ. Основы

- ***Корреляция применяется*** при изучении экспериментальных данных, представляющих собой измеренные значения двух признаков**.**
- В результате анализа **статистических данных** или **организованного эксперимента** регистрируются различные значения случайных величин входных факторов - Xij и выходной - Yi в каждом из опытов, (где i,j принимает значения натуральных чисел, i - в пределах от =1 до m, j пределах от 1 до n,) имеющие некоторую взаимосвязь между последовательностями значений наблюдаемых величин.
- При этом корреляционную зависимость между признаками можно **описывать разными способами**: соответствие между аргументом и функцией может быть задано таблицей, формулой, графиком и т. д.

Корреляционный анализ. Основы

- Совокупность точек на плоскости создает общую картину регрессии и позволяет построить некоторую **усредненную кривую взаимосвязи** параметров**.**
- **Корреляционной связью** между случайными переменными величинами называется функциональная связь между средним арифметическим значений наблюдений за выходной переменной, соответствующих данному значению входной. Характер и выраженность такой связи устанавливают с коэффициента корреляции, предложенног Браве.
- Наиболее простой вариант корреляционной связи, описывается коэффициентом парной корреляции, предложенным **Карлом Пирсоном 1884 году** математик, биолог, филос позитивист):



Корреляционный анализ. Основы

$$r_{xy} = \sum_{i=1}^{n} \left[\left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right) \right] / \delta_x \cdot \delta_y \left(n - 1 \right)$$

- где x_i, y_i – значения наблюдаемых входного и выходного параметров в i-том наблюдении;

  $\bar{x}$, $\bar{y}$ - средние значения параметров x и y наблюдений;

  n - общее число наблюдений;

  δx, δy - среднеквадратичные отклонения параметров x, y.

Среднеквадратичное отклонение параметра x определяется по известной формуле

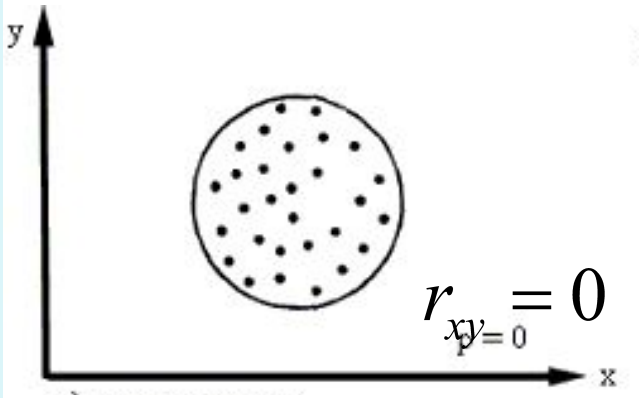
$$\delta_x = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x}) / (n - 1)},$$

Корреляционный анализ. Основы

- **Задача корреляционного анализа сводится к установлению направления и формы связи между признаками, измерению ее тесноты и к оценке достоверности выборочных показателей корреляции.**
  Корреляционная связь между признаками может быть *линейной и криволинейной (нелинейной), положительной и отрицательной.*
  - Величина коэффициента корреляции всегда заключена в пределах

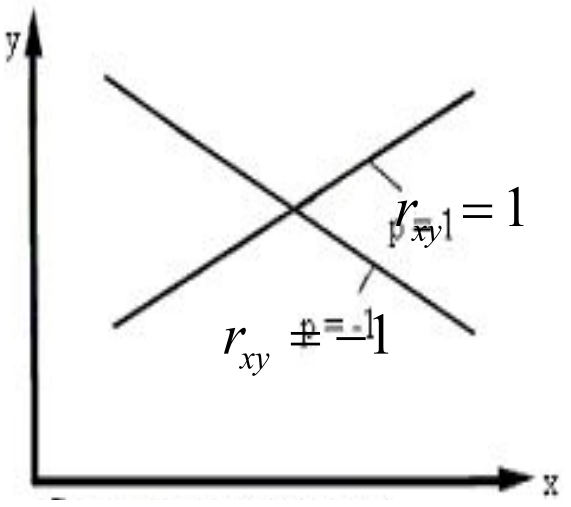
$$-1 \le r_{xy} \le 1$$

- Если коэффициент корреляции равен 0 то **корреляция отсутствует** – исследуемые параметры *x* и *y* не связаны линейной зависимостью, и являются независимыми случайными величинами, но это не свидетельствует об отсутствии связи – она может быть нелинейной.



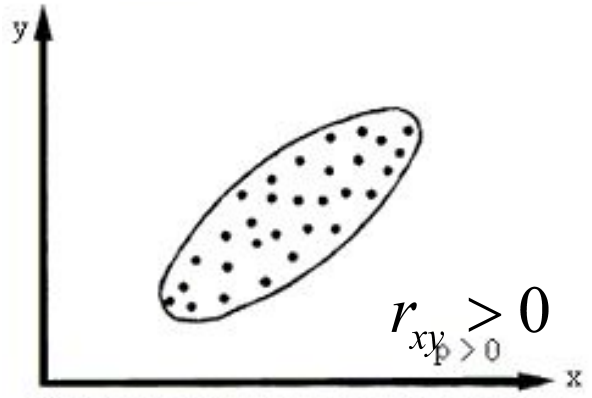
Корреляционный анализ. Основы

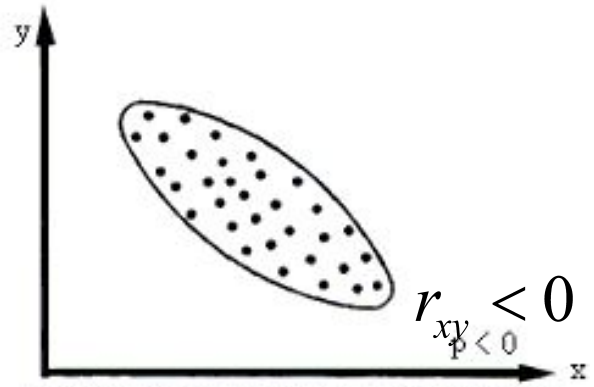
- **Если** коэффициент корреляции принимает максимально возможные значения **1** или **-1** то между случайными величинами X и Y существует линейная функциональная зависимость *(Y = a + вX)*. В этом случае говорят о **полной корреляции**.
- Если $r_{xy} = 1$, то значения определяют точки, лежащие на прямой линии, имеющей **положительный уклон** – с увеличением аргумента функция увеличивается.
- Если $r_{xy} = -1$, то значения определяют точки, лежащие на прямой линии, имеющей **отрицательный уклон** – с увеличением аргумента функция уменьшается.
- *графическая интерпретация на рисунке.*



Корреляционный анализ. Основы

- **Если** В промежуточных случаях $-1 < r_{xy} < 1$ точки попадают в область, ограниченную некоторым эллипсом. При этом, чем ближе расчетная величина к максимально возможной, тем уже эллипс и теснее экспериментальные значения группируются возле линии.
- **Прямая корреляция** отражает однотипность в изменении признаков: с увеличением (уменьшением) значений первого признака увеличиваются (уменьшается) значения и другого. **Обратная корреляция** указывает на увеличение первого признака при уменьшении второго или уменьшение первого признака при увеличении второго.





Корреляционный анализ. Основы

- Только по величине коэффициентов корреляции нельзя судить о достоверности корреляционной связи между признаками. Этот параметр зависит от числа степеней свободы k = n –2, где: n – число коррелируемых пар показателей X и Y. **Чем больше n**, тем выше **достоверность** связи при одном и том же значении коэффициента корреляции. **В практической деятельности**, когда число коррелируемых пар признаков X и Y не велико (n<=30), то при оценке зависимости между показателями используется следующую градацию:

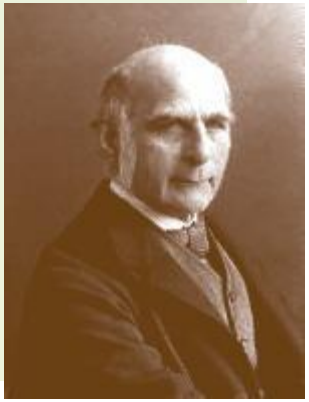
1) *высокая* степень взаимосвязи – значения коэффициента корреляции находится в пределах от 0,7 до 0,99;

2) *средняя* степень взаимосвязи – значения коэффициента корреляции находится в пределах от 0,5 до 0,69;

3) *слабая* степень взаимосвязи – значения коэффициента корреляции находится от 0,2 до 0,49.

Регрессионный анализ. Понятие

- В практических исследованиях возникает необходимость **аппроксимировать** (описать приблизительно) диаграмму рассеяния математическим уравнением. То есть зависимость между переменными величинами Y и X можно выразить аналитически с помощью формул и уравнений и графически в виде точек в системе прямоугольных координат.
- График корреляционной зависимости строится по уравнениям функции $\bar{y}_x = f(x)$, которые называются регрессией.
- Термин «**регрессия**» происходит от лат. regressio — движение назад. В нашем случае — **статистическая зависимость** среднего значения случайной величины от значений другой случайной величины или нескольких случайных величин; введена Фрэнсисом Гальтоном (английский статистик, психолог и антрополог; 1886).



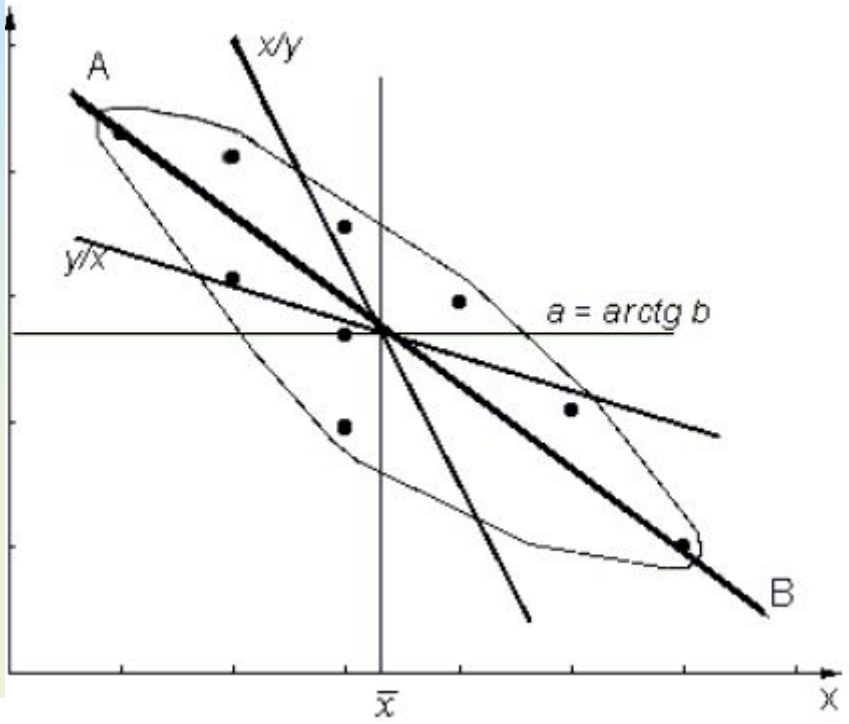
Уравнение линейной регрессии

- Обычно **признак** Y рассматривается как **функция** многих **аргументов** — x1, x2, x3, …— и может быть записана в виде:
  y = a + bx1 + cx2 + dx3 + … , *где: a, b, c и d — параметры уравнения, определяющие соотношение между аргументами и функцией.*

  В практике учитываются не все, а лишь некоторые аргументы. В простейшем случае, как при описании линейной регрессии, всего один **y = a + bx**
- *В уравнении параметр a — свободный член; графически он представляет отрезок ординаты (y) в системе прямоугольных координат. Параметр b называется коэффициентом регрессии. С точки зрения аналитической геометрии b— угловой коэффициент, определяющий наклон линии регрессии по отношению к осям, координат.*

Уравнение линейной регрессии

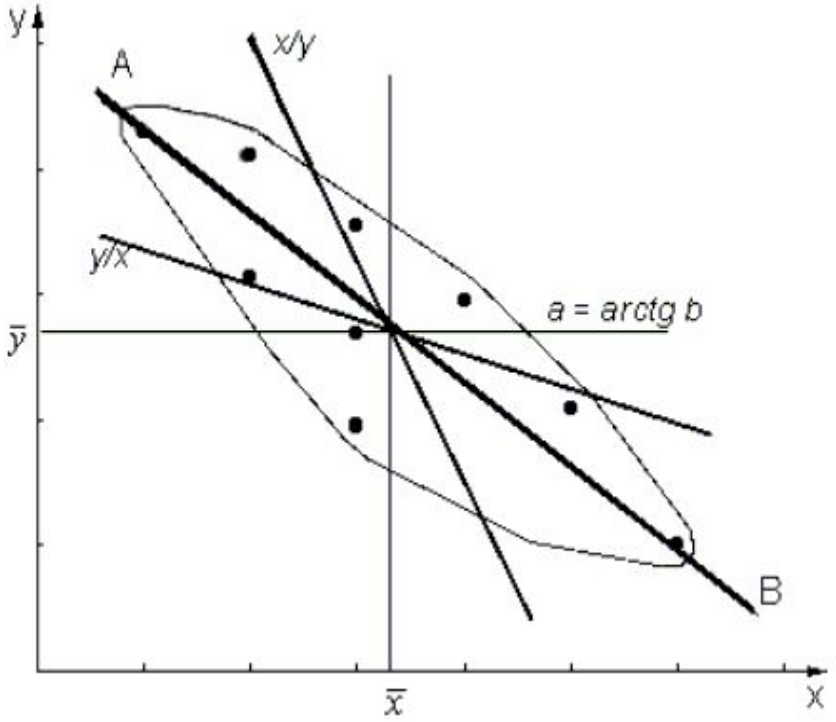
- В области регрессионного анализа **угловой коэффициент** показывает, насколько в среднем величина одного признака (Y) изменяется при изменении на единицу меры другого корреляционно признака X. Наглядное представление дают линий регрессии Y по X и X по Y в системе прямоугольных координат (см. рис.)



- Линии регрессии пересекаются в точке O(x,y) средних арифметических значений корреляционно связанных друг с другом признаков Y и X. Линия AB, проходящая через эту точку, изображает функциональную зависимость между переменными величинами Y и X, когда коэффициент корреляции $r_{xy} = 1$.

Уравнение линейной регрессии

- Чем **сильнее связь** между Y и X, тем ближе линии регрессии к АВ, и, наоборот, чем слабее связь между варьирующими признаками, тем **более удаленными** оказываются линии регрессии от АВ. При отсутствии связи между признаками, когда $r_{xy} = 0$, линии регрессии оказываются под прямым углом (90°) по отношению друг к другу.



- Уравнение регрессии тем лучше описывает зависимость, чем меньше рассеяние диаграммы, чем больше теснота взаимосвязи. Уравнение прямой линии пригодно для описания только линейных зависимостей. В случае нелинейных зависимостей математическая запись может отображаться уравнениями параболы, гиперболы и др.

Основы теории регрессионного анализа

- После **выбора гипотезы** о виде зависимости между случайными величинами (**вид уравнения**), которым описывается модель статистической связи, появляется необходимость **нахождения параметров** этого уравнения (свободного члена и коэффициентов).
- Эта задача решается с помощью регрессионного анализа. В общем случае из условия максимального приближения предполагаемой линии регрессии к точкам, отражающим опытные данные получается **система нормальных уравнений**. Для случая, когда все наблюдаемые значения за переменными x и y лежат точно на прямой линии, выполняется равенство:
  - $y_i - a_o - a_i x_i = 0,$

Основы теории регрессионного анализа

- На **практике** это равенство нарушается и для отдельных наблюдений появляется ошибка δ_i. Она определяется разностью между **измеренной** и **вычисляемой** по уравнению регрессии значениями переменной y в i – ом опыте. *Возникает задача нахождения коэффициентов уравнения, обеспечивающих минимальную ошибку* δ_i.
- Теория вероятностей показывает, что лучшим приближением будет такая линия, для которой **сумма** квадратов расстояний от точек до кривой будет минимальной. Этот метод называется **методом наименьших** квадратов, разработан **Гауссом** и называется принципом выравнивания; критерием выравнивания

$$\sum_{i=1}^{n} \delta_i^2 = Q_{\min}(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n} \left| f(x_i, a_0, \ldots, a_m) - y_i^2 \right|.$$

Основы теории регрессионного анализа

- Если погрешности δ_i подчиняются нормальному закону распределения, минимум можно найти, приняв к нулю частные производные по всем неизвестным:
  - $\mathbf{Q/ a_o = ...= Q/ a_m = 0}$
- После преобразования получается **система нормальных уравнений**. Решение этой системы позволяет найти искомые коэффициенты $(a_o, a_1,...,a_m)$ регрессии.

- **Иоганн Карл Фри́дрих Га́усс** немецкий математик, астроном и физик и философ. Считается величайшим математиков всех времён, «королём математиков». В 1794—95 годах осуществил первое применение к решению системы нормальных уравнений.

Условия применения метода наименьших квадратов

Входные факторы должны быть измерены с более высокой точностью в сравнении с ошибкой измерения выходной величины.

Некоррелированность входных факторов.

Измерения выходной переменной y должны представлять собой независимые друг от друга, нормально распределенные случайные величины с $D(\delta) = Const$.

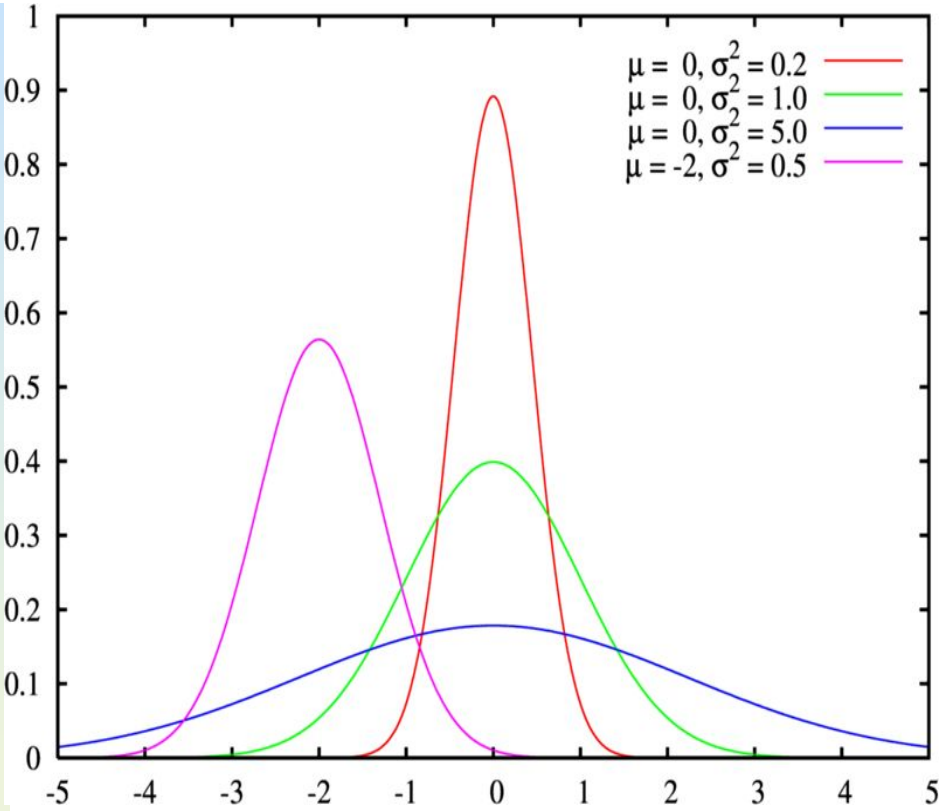
Вид зависимости должен быть известен.

Нормальное распределение Гаусса

- **Нормальное распределение**, также называемое гауссовым распределением, гауссианой или распределением Гаусса — распределение вероятностей, которое задается функцией плотности распределения

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

- Физическая величина, подверженная **влиянию** значительного числа **независимых** факторов, способных вносить с равной погрешностью **положительные** и **отрицательные** отклонения



Оценка качества полученной модели

- Мерой **степени соответствия** аппроксимирующей регрессии имеющимся значениям y_i является **коэффициент множественной корреляции** (или детерминации):

$$R^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

- Скорректированный коэффициент множественной корреляции

$$R_m^2 = \left[1 - (1 - R^2)\right](N-1)/(N-m-1),$$

- *Коэффициенты изменяются в пределах 0…1; чем* ***больше*** *его значение, тем выше качество модели.*
- Еще одним критерием качества модели является статистика J_m

$$J_m = (N + m + 1) \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (N - m - 1).$$

- *Качество полученной модели характеризуют* ***минимальные*** *значения* J_m

СПАСИБО ЗА ВНИМАНИЕ!

- Пожелания и предложения можно высказывать:
- лично - аудитория А 205;
- или письменно - timoshek@rgups.ru