# Лекция 3: Регрессионный, корреляционный и дисперсионный виды анализа

1. *Регрессионный анализ.*
2. *Корреляционный анализ.*
3. *Дисперсионный анализ.*

# 1 Регрессионный анализ

Функциональная зависимость может быть представлена в виде «ящика»: он преобразует вход $X = \{x_1, x_2, \ldots, x_N\}$ к выходу $Y = \{y_1, y_2, \ldots, y_N\}$.

Функция ящика: одномерная («один вход» - «один выход»), или многомерная.

что известно об объекте:

**все** **структура** **ничего**

| | белый | серый | черный |
|---|---|---|---|
| структура | + | + | - |
| колич. значения параметров | + | - | - |

*Задача* ***регрессионного анализа*** *– нахождение уравнения зависимости откликов от фактора, т.е. восстановление функциональной зависимости параметров по данным эксперимента.*

*Искомое уравнение –* ***уравнение (функция) регрессии****.*

Рассмотрим линейную одномерную регрессию (один вход – один выход).

Экспериментальные точки могут быть представлены на декартовой плоскости (диаграмма рассеяния). Они выстраиваются почти в прямую линию.

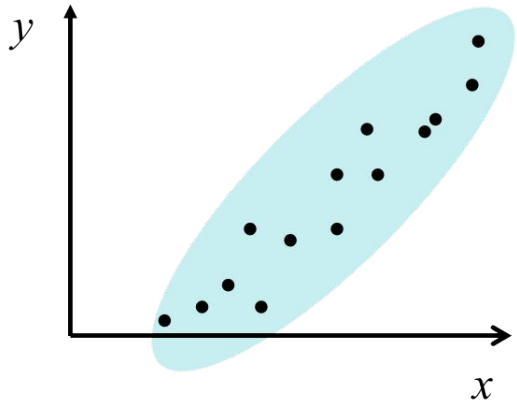
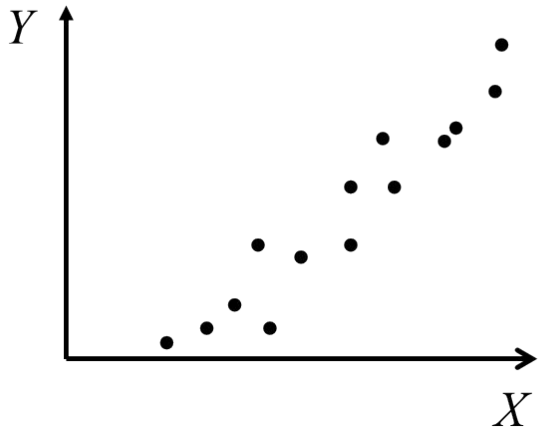


диаграмма рассеяния

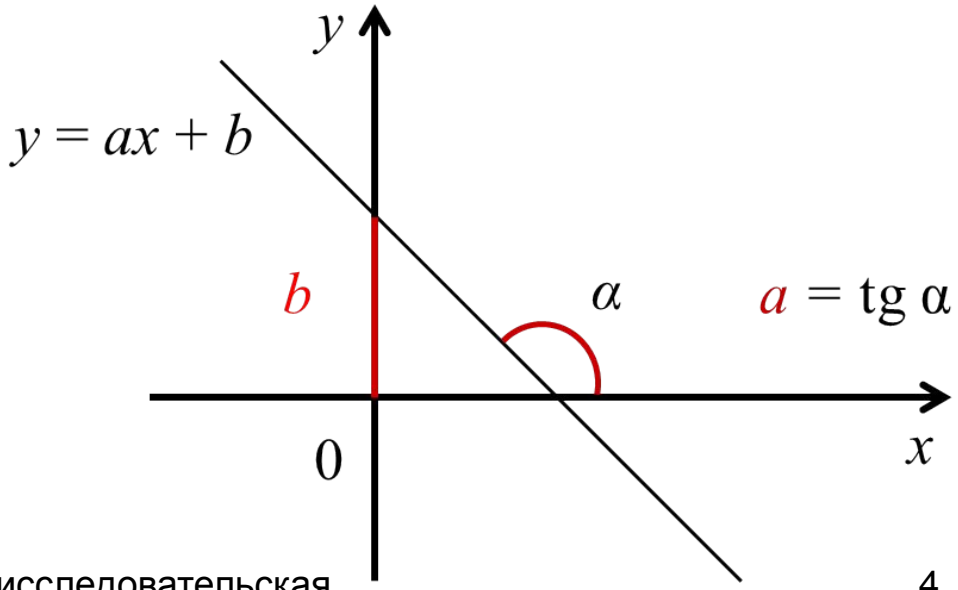
# Алгоритм нахождения одномерной линейной функции регрессии



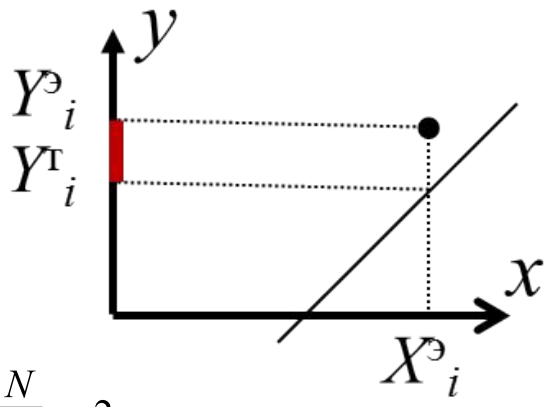

0. Предварительная оценка линейности по ***диаграмме рассеяния*** - *отображение данных X и Y в виде точек на декартовой плоскости ($X_i$, $Y_i$).*

1. Выдвижение $H_0$: функция регрессии («черного ящика») имеет вид

$$Y = f(X) = aX + b$$



2. Для каждой точки находится разность $\varepsilon_i$ между экспериментальным значением отклика $Y_i$ и «теоретическим» значением отклика $Y^T_i$

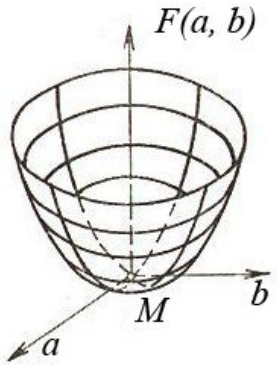



$$\varepsilon_i = Y_i - Y_i^T = Y_i - (aX_i + b)$$

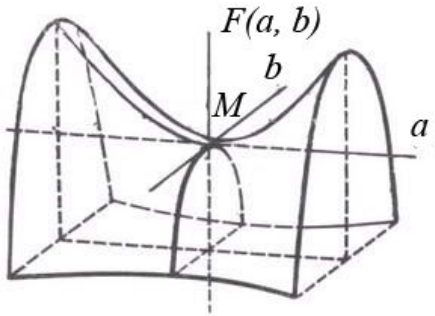
3. Находится суммарная ошибка $F(a, b) = \sum_{i=1}^{N} \varepsilon_i^{\,2} =$

$$= \left(\sum_{i=1}^{N} X_i^{\,2}\right) a^2 + Nb^2 + \left(2\sum_{i=1}^{N} X_i\right) ab - \left(2\sum_{i=1}^{N} X_i Y_i\right) a - \left(2\sum_{i=1}^{N} Y_i\right) b + \sum_{i=1}^{N} Y_i^2$$

$F(a, b)$ – квадратичная, ***a*** и ***b*** – неизвестные.




эллиптический параболоид: есть extr




гиперболический параболоид: нет extr, только седловая точка

Для нахождения min $F(\boldsymbol{a}, \boldsymbol{b})$

а) необходимые условия экстремума => находим координаты *a, b* т.н. стационарной точки *M*:

$$\begin{cases} \dfrac{\partial F}{\partial a} = 0, \\ \dfrac{\partial F}{\partial b} = 0. \end{cases}$$

$$a = \frac{N\left(\sum_{i=1}^{N} X_i Y_i\right) - \left(\sum_{i=1}^{N} X_i\right)\left(\sum_{i=1}^{N} Y_i\right)}{N\left(\sum_{i=1}^{N} X_i^{\,2}\right) - \left(\sum_{i=1}^{N} X_i\right)^2} \qquad b = \frac{\left(\sum_{i=1}^{N} X_i^{\,2}\right)\left(\sum_{i=1}^{N} Y_i\right) - \left(\sum_{i=1}^{N} X_i Y_i\right)\left(\sum_{i=1}^{N} X_i\right)}{N\left(\sum_{i=1}^{N} X_i^{\,2}\right) - \left(\sum_{i=1}^{N} X_i\right)^2}$$

б) достаточные условия экстремума => проверка того, что точка с координатами (*a, b*) – минимум функции.

$$A = \left.\frac{\partial^2 F}{\partial a^2}\right|_M \qquad B = \left.\frac{\partial^2 F}{\partial a\,\partial b}\right|_M \qquad C = \left.\frac{\partial^2 F}{\partial b^2}\right|_M \qquad D = AC - B^2$$

В нашем случае

$$A = \frac{\partial^2 F}{\partial a^2} = 2\left(\sum_{i=1}^{N} X_i{}^2\right) \quad B = \frac{\partial^2 F}{\partial a\,\partial b} = 2\left(\sum_{i=1}^{N} X_i\right) \quad C = \frac{\partial^2 F}{\partial b^2} = 2N$$

$$D = AC - B^2 = 4N\left(\sum_{i=1}^{N} X_i{}^2\right) - 4\left(\sum_{i=1}^{N} X_i\right)^2$$

Если $D<0$ F(a, b) – гиперболический парабалоид.

Если $D>0$ F(a, b) – эллиптический парабалоид:

- $A>0$ в ($a$, $b$) – min;
- $A<0$ в ($a$, $b$) – max.

Для вычисления $a$ и $b$ можно использовать выражения:

$$a = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - \overline{X}^2} \qquad b = \frac{\overline{Y}\,\overline{X^2} - \overline{X}\cdot\overline{XY}}{\overline{X^2} - \overline{X}^2}$$

# Адекватность регрессионной модели

Выборочный коэффициент детерминации $R^2$

$$R^2 = \frac{\sum (Y_i^T - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

→ объясненные моделью отличия

→ общее отклонение

$$R^2 \in [0,1]$$

$R^2 \approx 1$ модель хорошего качества.
$R^2 \approx 0$, построенная модель плохого качества.

На $(R^2)\cdot 100\%$ найденная функция регрессии описывает связь между исходными значениями $Y$ и $X$;

$(1-R^2)\cdot 100\%$ отклонения значений $Y$ обусловлены факторами, не включенными в регрессионную модель.

Если $R^2 \geq 0{,}75$, по модели можно делать прогноз значений в пределах исходного диапазона данных.

## *Алгоритм оценки адекватности:*

1 **$H_0$**: ***генеральное значение $R^2$ незначимо***.
Т.е даже если рассчитанное (выборочное) значение $R^2$ близко к 1, это получилось только из-за выборки.

2 Статистика критерия:

$$F_{\text{набл}} = \frac{R^2}{1-R^2} \cdot \frac{N-p-1}{p}$$

3 Задаемся уровнем значимости (α=0,05)

4 Находим $F_{\text{кр}}$ – значение критерия Фишера для заданного уровня значимости α с числом степеней свободы $k_1=p$, $k_2=N-p-1$ (для линейной регрессии $p=1$).

5 Если $F_{\text{набл}} \le F_{\text{кр}}$, **$H_0$** принимается (модель неадекватна).

# 2 Корреляционный анализ

Рассмотрим полученные в ходе эксперимента наборы данных: $X = \{x_1, x_2, \ldots, x_N\}$ $Y = \{y_1, y_2, \ldots, y_N\}$

*Задача **корреляционного анализа** – обнаружение взаимосвязи между двумя параметрами и количественная оценка степени неслучайности их совместного изменения.*

Исследуемые величины могут быть как двумя показателями в одной выборке, так и двумя различными выборками.

выборка →

параметры

| **Человек** | **1** | **2** | **3** | **…** | ***n*** |
|---|---|---|---|---|---|
| **Рост** | 178 | 150 | 167 | … | 166 |
| **Вес** | 60 | 55 | 60 | … | 59 |

параметр →

| **Пары близнецов** | | **1** | **2** | **3** | **…** | ***n*** |
|---|---|---|---|---|---|---|
| **Усидчивость (1-10)** | Близнец1 | 6 | 4 | 7 | … | 7 |
| | Близнец2 | 6 | 8 | 9 | … | 1 |

выборки

Если есть связь между величинами, корреляционный анализ показывает:

- растет/уменьшается один параметр с ростом другого;
- насколько сильно один показатель влияет на другой.

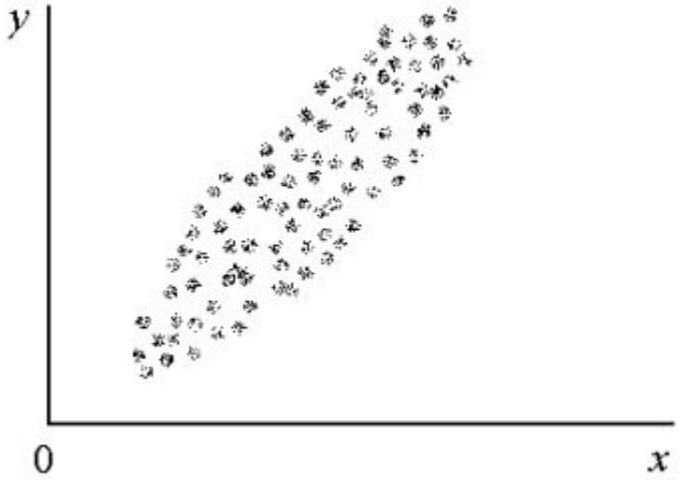
Корреляционный анализ помогает установить возможность предсказания вероятных значений одного показателя с помощью известных значений другого.

Изображение исходных данных - ***корреляционное поле***:

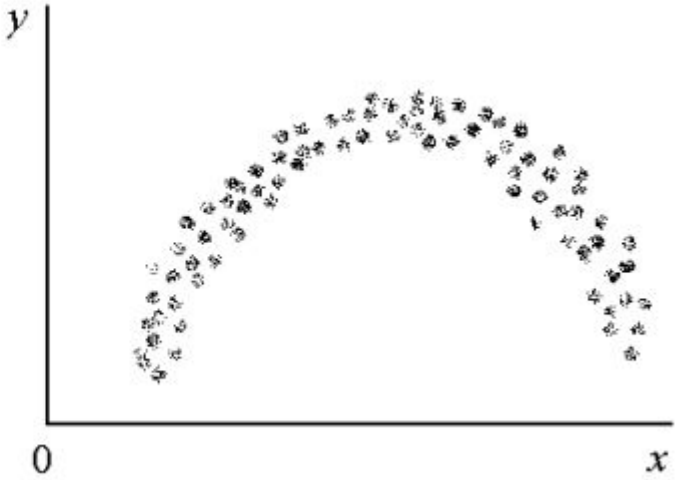
- по оси абсцисс шкала для одного показателя (выборки);
- по оси ординат шкала для другого показателя (выборки).

По расположению точек на корреляционном поле можно судить о наличии/отсутствии связи, ее силе и характере.

линейная




нелинейная




Для определения взаимосвязи между параметрами используется ***коэффициент корреляции*** – только для случая линейной взаимосвязи между параметрами (для нелинейной связи дает ложные значения).

Классификация по силе связи:

- функциональная – есть жесткая зависимость между двумя параметрами, которую можно записать в виде функции без сглаживания;
- сильная;
- умеренная;
- слабая;
- отсутствующая – связи нет.

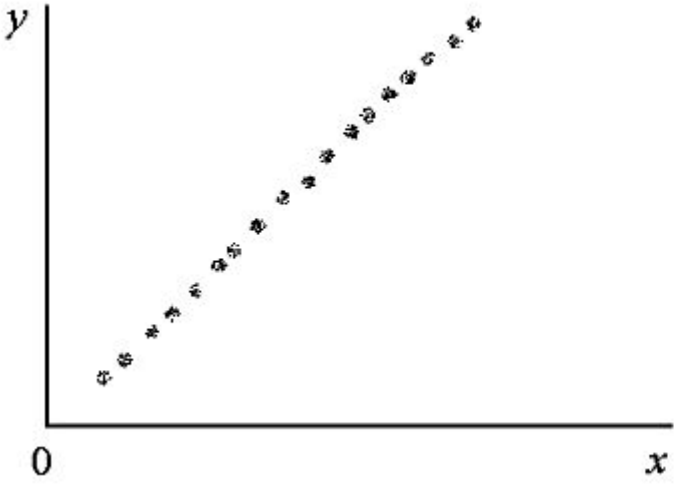
Классификация по направлению связи:

- положительная, характеризующая прямую зависимость между параметрами, когда увеличение одного параметра приводит к увеличению другого;
- отрицательная, характеризующая обратную зависимость между параметрами, когда увеличение одного параметра приводит к уменьшению другого.

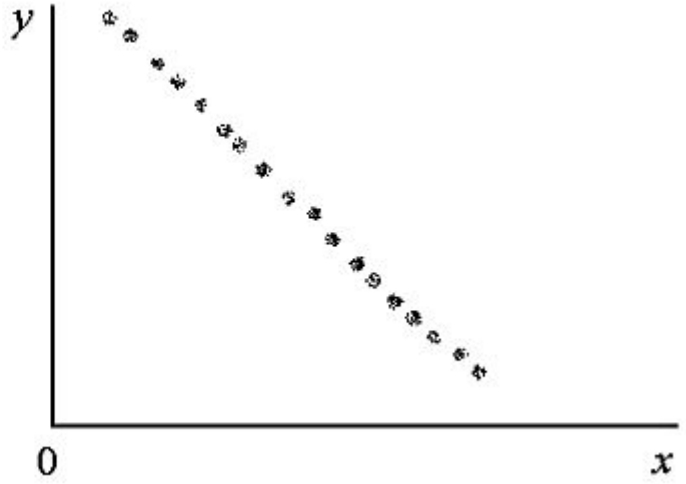
# *Классификация связей по силе и направлению на корреляционном поле*

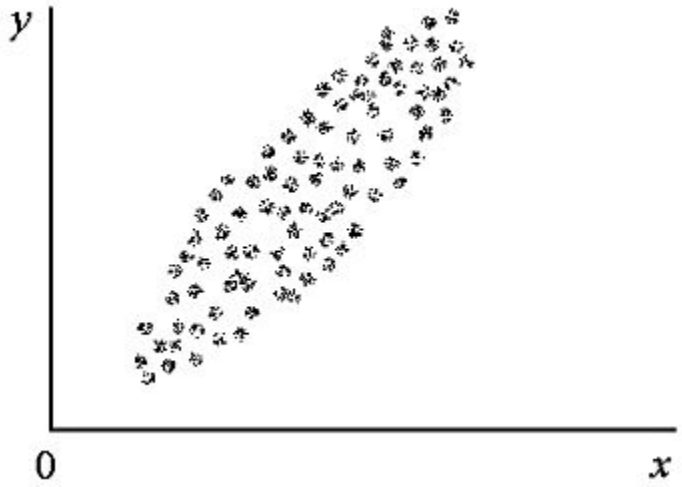
положительная

отрицательная

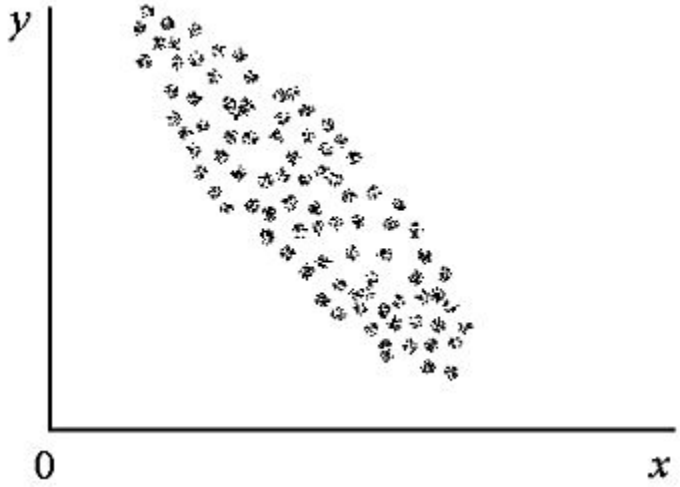



функциональные

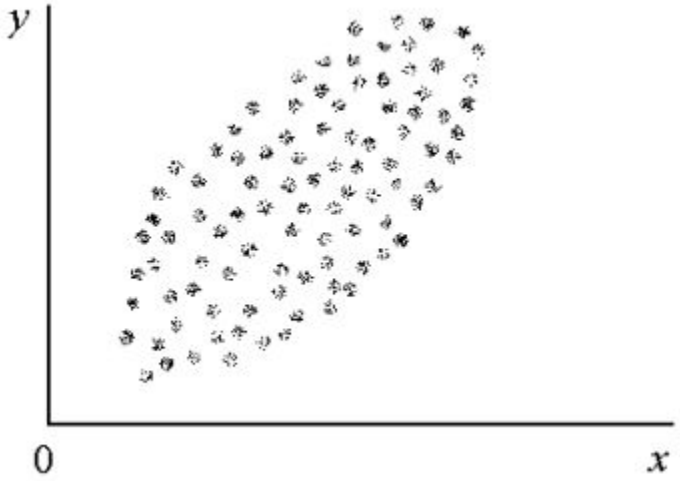



сильные

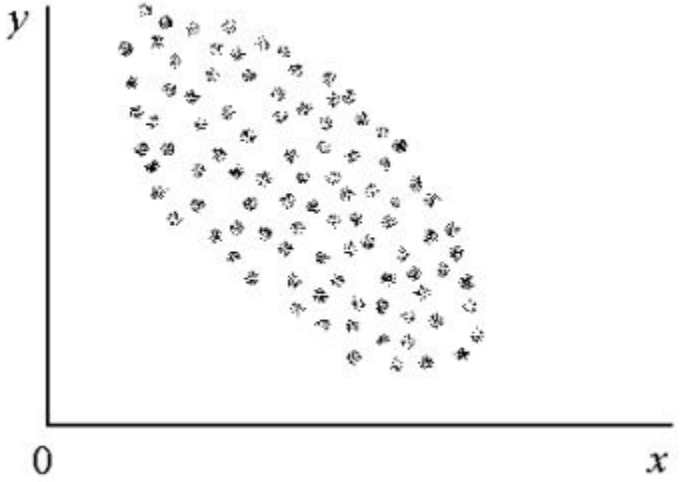


положительная

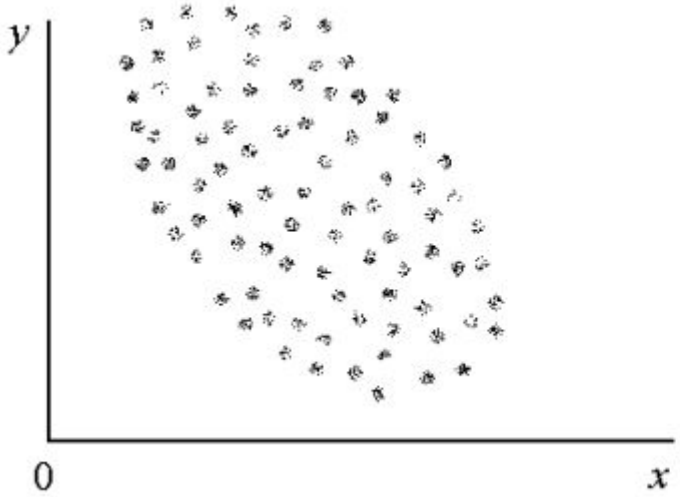



умеренные

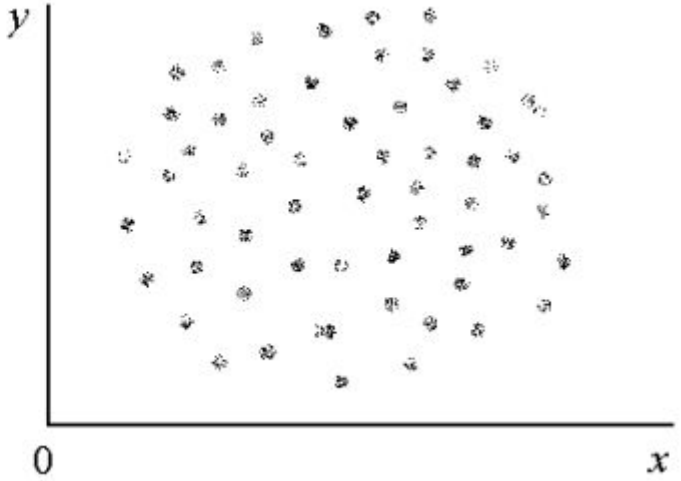
отрицательная




отрицательная слабая




отсутствует



***Коэффициент линейной корреляции:***

Пусть есть случайные векторы $X=\{x_i\}$, $Y=\{y_i\}$, $i=1\ldots N$:

$$r = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_X \sigma_Y} \qquad \text{или} \qquad r = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{Y})^2}}$$

$$\overline{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i y_i)$$

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad \sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N}} \qquad \sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \overline{Y})^2}{N}}$$

Для малых объемов выборки ($N \leq 100$) корректировка:

$$r' = r\left(1 + \frac{1 - r^2}{2(N-3)}\right)$$

Значения коэффициента корреляции: $-1 \leq r \leq 1$

- знак определяет характер связи (положительная или отрицательная)
- модуль – силу связи.

При $r = 0$ связь отсутствует, т.е. изменение $X$ не приводит к изменению $Y$.

При $|r| = 1$ наблюдается строгая функциональная зависимость (т.е. есть функция $Y=f(X)$).

При $|r| \to 0$ зависимость одной переменной от другой все больше уменьшается, то есть «облако» значений на корреляционной плоскости становится шире и все более округлым.

При $|r| \to 1$ «облако» значений «концентрируется» в график функции зависимости.

# Сила связи между параметрами в зависимости от величины $r$

| Значение $r$ | Сила связи |
|---|---|
| $\lvert r \rvert = 1$ | функциональная |
| $0{,}7 \le \lvert r \rvert < 1$ | сильная |
| $0{,}5 \le \lvert r \rvert \le 0{,}7$ | умеренная |
| $0{,}3 \le \lvert r \rvert \le 0{,}5$ | слабая |
| $0 < \lvert r \rvert \le 0{,}3$ | практически отсутствует |
| $\lvert r \rvert = 0$ | отсутствует |

линейная регрессия $Y$ на $X$: $$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right)$$

уравнение линейной регрессии $$Y = r\frac{\sigma_Y}{\sigma_X}X + \left(\overline{Y} - r\frac{\sigma_Y}{\sigma_X}\overline{X}\right)$$

## *Значимость генерального коэффициента линейной корреляции:*

Выборочный коэффициент $r$ – оценка генерального коэффициента корреляции, который показывает реальную связь между $X$ и $Y$.

Из-за конечного размера выборок возможен случай, когда выборочный $r \approx 1$, а генеральный $r \approx 0$. Т.е. выборочный коэффициент корреляции покажет отсутствующую (нулевую) на генеральной совокупности сильную связь между параметрами.

Доказательство значимости проводится методом проверки статистических гипотез.

1 Выдвигаются нулевая и альтернативная гипотезы:

- нулевая - о равенстве нулю генерального коэффициента корреляции $H_0$: $r_s=0$
- альтернатива – $H_1$: $r_s\neq 0$

2 Задается уровень значимости $\alpha=0,05$.

3 Вычисляется статистика

- для $N\geq 100$

$$t_{\text{набл}} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{N-2}$$

- для N<100

$$t_{\text{набл}} = 0,5\ln\left(\frac{1+r'}{1-r'}\right)\sqrt{N-3}$$

4 Находится $t_{\text{кр}}$ – значение коэффициента Стьюдента $t(P=1-\alpha,\ \propto)$

5 Если $t_{\text{набл}} > t_{\text{кр}}$ , то $H_0$ отвергается, т.е. генеральный коэффициент корреляции значимо больше нуля.

## *Значимость различия между двумя коэффициентами линейной корреляции:*

Значение $r$ может меняться в зависимости от объема выборки или самих значений. Если есть две пары выборок, принадлежат ли они одной генеральной совокупности?

Пусть есть выборки

- $X1=\{x1_i\}$, $Y1=\{y1_i\}$, $i=1\ldots N$, с выборочным $r1$;
- $X2=\{x2_j\}$, $Y2=\{y2_j\}$, $j=1\ldots M$, $M\neq N$ с выборочным $r2$;
- $r1\neq r2$.

Имеют ли эти выборки общий генеральный коэффициент линейной корреляции?

Доказательство методом проверки статистических гипотез.

1 Выдвигаются нулевая и альтернативная гипотезы:

- нулевая - о незначимости различий между двумя генеральными коэффициентами линейной корреляции $H_0: r1_s = r2_s = r_s$
- альтернатива – $H_1: r1_s \neq r2_s$

2 Задается уровень значимости $\alpha = 0{,}05$.

3 Вычисляется статистика

$$t_{\text{набл}} = 0{,}5 \ln\left(\frac{(1+r1)(1-r2)}{(1-r1)(1+r2)}\right) \frac{1}{\sqrt{\frac{1}{N-3} + \frac{1}{M-3}}}$$

4 Находится $t_{\text{кр}}$ – значение коэффициента Стьюдента $t(P=1-\alpha, \propto)$

5 Если $t_{\text{набл}} > t_{\text{кр}}$ , то $H_0$ отвергается, т.е. нельзя считать, что обе пары взяты из одной генеральной совокупности.

# 3 Дисперсионный анализ (ANOVA)

рассматривает результаты наблюдений, которые зависят от одновременно действующих факторов.

Результат:

- нахождение наиболее значимых факторов;
- оценка влияния факторов на исследуемый процесс.

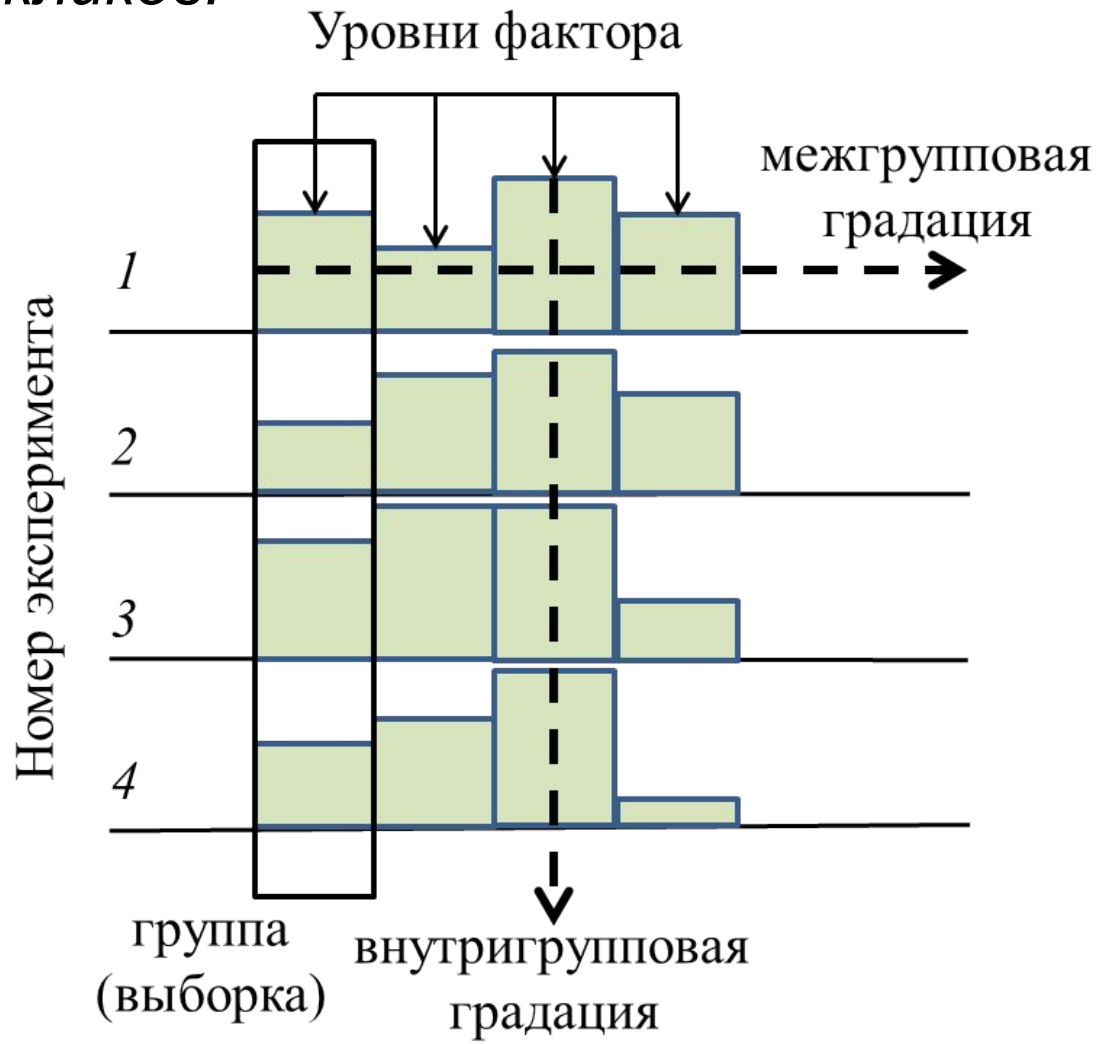
***Суть анализа***: разделение общей дисперсии на отдельные компоненты, обусловленные влиянием факторов, и проверке гипотез о значимости влияния факторов на среднее значение наблюдаемой величины.

Предположения:

- распределение исходных случайных величин нормально;
- дисперсии данных одинаковы для экспериментов, выполненных на различных уровнях изучаемого фактора.

***Группа*** *- набор значений откликов, полученных при фиксированных уровнях факторов.*
*Градация - изменение откликов:*

- ***межгрупповая градация*** *– изменение откликов, соответствующее уровням факторов;*
- ***внутригрупповая градация*** *– изменение откликов внутри одной выборки, соответствующей одному уровню факторов.*



Пусть есть $m$ выборок $x_1,...,x_m$ одинакового объема $n$.

Исходные данные могут быть представлены в виде статистической таблицы:

| Номер эксперимента | Уровни фактора *Fact* | | | | |
|---|---|---|---|---|---|
| | $F_1$ | … | $F_j$ | … | $F_m$ |
| 1 | $x_{11}$ | … | $x_{1j}$ | … | $x_{1m}$ |
| … | … | | … | | … |
| $i$ | $x_{i1}$ | … | $x_{ij}$ | … | $x_{im}$ |
| … | … | | … | | … |
| $n$ | $x_{n1}$ | … | $x_{nj}$ | … | $x_{nm}$ |

В процессе анализа рассчитываются дисперсии:

- общая (дисперсия комплекса);
- межгрупповая (факторная);
- внутригрупповая (остаточная).

Алгоритм одномерного однофакторного ДА

1 Задается уровень значимости $\alpha=0,05$.

2 Гипотеза: $H_0 : Mx_1 = Mx_2 = Mx_3 = ... = Mx_m$

3 Расчет средних:

- внутригрупповое

$$\bar{x}_{*j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$$

- межгрупповое

$$\bar{x}_{i*} = \frac{1}{m}\sum_{j=1}^{m} x_{i,j}$$

- общее

$$\bar{x}_{**} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}$$

| Номер эксперимента | Уровни фактора *Fact* | | | | | |
|---|---|---|---|---|---|---|
| | $F_1$ | … | $F_j$ | … | $F_m$ | $\Sigma/m$ |
| 1 | $x_{11}$ | … | $x_{1j}$ | … | $x_{1m}$ | $\bar{x}_{1*}$ |
| … | … | | … | | … | … |
| $i$ | $x_{i1}$ | … | $x_{ij}$ | … | $x_{im}$ | $\bar{x}_{i*}$ |
| … | … | | … | | … | … |
| $n$ | $x_{n1}$ | … | $x_{nj}$ | … | $x_{nm}$ | $\bar{x}_{n*}$ |
| $\Sigma/n$ | $\bar{x}_{*1}$ | | $\bar{x}_{*j}$ | | $\bar{x}_{*m}$ | $\bar{x}_{**}$ общее среднее |

*межгрупповые средние*

*внутригрупповые средние*

4 Расчет сумм квадратов отклонений:

- общая сумма квадратов отклонений от общего среднего

$$R_{\text{общ}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_{**}\right)^2$$

- факторная сумма квадратов отклонений групповых средних от общего среднего (межгрупповое рассеяние)

$$R_{\text{факт}} = n\sum_{j=1}^{m}\left(\bar{x}_{*j} - \bar{x}_{**}\right)^2$$

различия между средними значениями в группах

- остаточная сумма квадратов отклонений (внутригрупповое рассеяние)

$$R_{\text{ост}} = R_{\text{общ}} - R_{\text{факт}}$$

не может быть предсказано или объяснено

5 Расчет несмещенных выборочных дисперсий:

- общая

$$S^2_{\text{общ}} = \frac{R_{\text{общ}}}{nm - 1}$$

- факторная

$$S^2_{\text{факт}} = \frac{R_{\text{факт}}}{m - 1}$$

- остаточная

$$S^2_{\text{ост}} = \frac{R_{\text{ост}}}{m(n-1)}$$

6 Расчет статистики:

$$F_{\text{набл}} = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}$$

6 Нахождение $F_{кр}$ по числу степеней свободы $f_1=m-1$, $f_2=m(n-1)$ и уровню значимости $\alpha$ (таблицы значений распределения Фишера)

7 Если $F_{набл} > F_{кр}$, гипотеза отвергается, т.е. фактор оказывает существенное влияние на параметр и его надо учитывать.

Если гипотеза принимается, фактор – несущественный, им можно пренебречь.

Иногда дисперсионный анализ применяется для доказательства того, что выборки однородны:

дисперсии одинаковы + математические ожидания одинаковы => выборки можно объединить в одну и получить более полную информацию.