**Первая лекция.**

Мат статистика решает задачи, являющиеся обратными задачам теории вероятности

В теории вероятности после задания и определения того или иного случайного явления требуется рассчитать вероятностные характеристики в рамках заданной модели. Из имеющихся результатов эксперимента, называющихся статистическими данными, требуется определить или уточнить математическую модель (феноменологическую) того или иного явления, его определяющую и рассчитать вероятность некоторого события.

Например, схема Бернули:
*- Кидаем монету,*
*- рождается мальчик / девочка*
[1]

Мат статистика это обратная задача, поэтому более сложная. Трактовок обратных задач в разы больше (утраивается число). Задачи мат. статистики предполагают:

1) Оценку неизвестных параметров по результатам экспериментов.
(В теории вероятности число успехов, в обратной задаче найти вероятность)
2) Интервальное оценивание неизвестных параметров, при этом требуется построить интервал, в которых попадает расчётное число параметров с заданной вероятность гамма. Параметр гамма называется **коэффициентом доверия**.
Детализация: 0.9, 0.95, 0.99, 0.995, 0.999 (значения гамма выбираются из этого ряда)
Гамма измеряется в единицах вероятностей. Например какова вероятность, что с гамма придёт такое-то число студентов на пару.
3) Задача проверки статистических гипотез. Требуется на основе результатов эксперимента проверить то или иное предположения.

Мат статистика, как и теория вероятностей занимается изучением мат моделей случайных явлений и процессов по данным статистического эксперимента. Модели, ориентированные на решения задачи теории вероятности называются вероятностными, а на решение задач мат статистики - статистическими.

Стохастические (случайные) модели разделяются на:
1. Статистические
2. Вероятностные
1. и 2. относятся к феноменологическим моделям (из жизни, из опытов) или к аналитическим (если доказано или обоснована, связь или взаимосвязь входа и выхода в задаче).

Выборочная терминология.
В мат. статистике часто используется выборочная терминология, основанная на урновой схеме. [3]

[3] - генеральная совокупность. [4] - выборка n малых из генеральной совокупности.

Выборка может производиться с возвращением и без возвращения.
Если выборка производится с возвращением, величины из [3] независимы. В этом случае говорят, что набор чисел из [4] называется независимой повторной случайно выборкой объема n из [3]. Терминология сохраняется и в случае бесконечной генеральной совокупности.

Если значения в [4] расположить в порядке возрастания или убывания: [5], в этом случае [5] называется вариационным рядом.

Эмперическая функция распределения построенная на основе выборки [4] или вариационного ряда [5] называется функция [6], в которой:
r(x) - дискретная функция, кол-во значений [7], например ряд отказа работы лампочек: (0, 0, 5, 7, 10, 11, 16,15, 43,49)
График, таблица будет такой: [8]. На основе таких графиков рассчитываются гарантии.

Формально эмпирическая функция распределения может быть представлена в виде вариационного ряда: [9], где:
j - кол-во чисел xi из [5] не превышающих (меньше или равно) текущего значения x (дискретная переменная).

Пусть F(x) - истинная или теоретическая функция распределения для данной генеральной совокупности, тогда по теореме Гливенко-Кантелли, при n, стремящемся к бесконечности с вероятностью 1 выполняется соотношение [10] *(заботать супремыму и инфинимумы)*, прирассмотрение sup рассматриваются крайние точки.

Величина [11] называется выборочным или эмперическим средним.
Величина [12] $\sigma = \frac{1}{n}\sum_{i=1}^{n} (x_i - \underline{x})^2$является аналогом дисперсии, является выборочной или эмпирической дисперсией.
Величина [13] $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \underline{x})^2}$ является выборочным или эмпирическим СКО (среднеквадратичное отклонение)

Дисперсия и СКО позволяют оценить размах выборки относительно выборочного среднего значения $\underline{x}$

Если выборка представлена вариационным рядом, то величина $R_n = |x_n - x_1|$— называется **размахом выборки [14]**

Пример: (15, 17, 23, 28, 31, 37, 39, 45) — вариационный ряд. Размах выборки = 30.

Если для построения статистического критерия или метода решения выборочного среднего и СКО или дисперсии недостаточны, вводят моменты более высокого порядка. В общем виде : **[16]** - $S^r = \frac{1}{n}\sum_{i=1}^{n} \quad (x_i - a)^r$— момент порядка r, где a - смещение. Если данных недостаточно - нужно доопределить систему, в этом случае и вводятся моменты более высокого порядка, порядка r.

Смещения считаются по формуле $x_i = u + y_i;\ i = \underline{1,n};\ \underline{x} = u + \frac{1}{n}\sum_{i=1}^{n} \quad y_i$(, где u является основанием)

Выборочные моменты для выборки с повторяющимися значениями:
$\sigma = \frac{1}{n}\sum_{i=1}^{k} \quad (m_i x_i - \underline{x})^2$ Например: (17, 17, 18, 18, 18, 19). $m_1 = 2;\ m_2 = 3;\ m_3 = 1$, где $\sum_{i=1}^{k} \quad m_i = n$

# Точечная оценка параметров

Пусть имеется некоторая случайная величина с функцией распределения:$F(x,\theta), f(x,\theta)$, где F — **функция распределения**, f — **плотность распределения**
$\theta$ - единственный параметр распределения или многомерный вектор (при этом отличается метод оценивания)
$x = (\xi_1, \xi_2, \ldots, \xi_n)$ : x — переменная. Конкретная её реализация — $\xi_i$

Здесь и далее для однозначности будем считать $\theta$ многомерным вектором, который нужно найти.
при различных возможных параметрах $\theta$, называется **параметрическим семейством распределения**.

Любая функция от значений выборки (результатов наблюдений) называется **статистикой эксперимента.** [20] (если зависит от всех результатов эксперимента)
Пример: $y = kx + l$

$$y_1 = 31$$
$$y_2 = 35$$
$$y_3 = 45$$
$$y_4 = 45$$
$$y_5 = 91$$

Если зависит от всех значений выборки $\gamma = \varphi(\xi_1, \xi_2, \ldots, \xi_n)$, где $\gamma$— **статистика эксперимента**.

В этом случае задача построения точечной оценки параметра $\theta$сводится к нахождению функции $\varphi$ от результатов наблюдений. Статистику можно выбрать любую. Поэтому выбираем вектор набора параметров $t$.

Если расшифровать $\theta$ как $\theta = (\theta_1, \theta_2, \ldots, \theta_1)$, где
$\theta_1 = \varphi_1(\xi_1, \xi_2, \ldots, \xi_n)$
$\theta_2 = \varphi_2(\xi_1, \xi_2, \ldots, \xi_n)$
$\theta_k = \varphi_k(\xi_1, \xi_2, \ldots, \xi_n)$

## Свойства оценок

**Оценка** $\theta\hat{}$ называется **несмещенной**, если её математическое ожидание совпадает со значением параметра $M\theta\hat{} = \theta$, где
M - математическое ожидание (в зарубежной лит-ре E)
**Оценка** $\theta\hat{}$ называется **асимптотически несмещенной**, если $\lim\limits_{n\to\infty} M\theta\hat{} = \theta$, где n — количество испытаний
Если оценка при $n \to \infty$по вероятности по истинному значению параметров, то **оценка** называется **состоятельной для** $\forall\epsilon > 0$при всех возможных значениях $\theta$: $P(|\theta\hat{}_n - \theta| > \epsilon) = 0$

Пусть $\theta\hat{}_n$есть асимптотически несмещенная оценка параметра $\theta$и $\lim\limits_{n\to\infty} M\theta\hat{} = \theta$и $\lim\limits_{n\to\infty} D(\theta\hat{}_n) = \theta$— оценка состоятельная, тогда можем остановить эксперимент.
$D(\theta\hat{}_n)$ определяет разброс $\theta\hat{}_n$относительно точного значения параметра $\theta$

Пусть имеются $\theta\hat{}_1$и $\theta\hat{}_2$ , тогда $D(\theta\hat{}_1) = M(\theta\hat{}_1 - \theta)^2 \le M_2(\theta\hat{}_2 - \theta)^2 = D(\theta\hat{}_2)$.
Поэтому $\theta\hat{}_1$лучше.

**Вторник (09.10.13)**

Пусть $x_1, x_2, \ldots, x_n$выборка объема n (независимая случайная повторная) из генеральной совокупности с заданной функцией распределения $F(x, \theta)$и плотностью $f(x, \theta)$
В качестве оценки неизвестного параметра берется то его значение, при котором точное (истинное/теоретическое) значение параметра первого момента совпадает с эмпирическим значением, найденным по результатам эксперимента.
$\mu_1 = \int_{-\infty}^{\infty} x\, f(x, \theta)\, dx$[21], — теоретическое или истинное знанчение,
$\mu_2 = \underline{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$— выборочное среднее для объема n генеральной совокупности],
$\int_{-\infty}^{\infty} x\, f(x, \theta)\, dx = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \theta\hat{}$— оценка.

Т.е. идея метода момента: приравняли теоретическое (истинное) значение для генеральной совокупности к выборочному значению объема n к этой же совокупности и оценели параметр $\theta\hat{}$

Функция распределения [22]

$$F(x, \lambda) = 1 - e^{-\lambda x} \Rightarrow f(x, \lambda) = e^{-\lambda x}$$

$$\int_{-\infty}^{\infty} xe^{-\lambda x}dx = c$$

Далее [23]

$$\theta = (\theta_1, \theta_2, \ldots, \theta_1)$$
$$\mu_1(\theta_1, \theta_2, \ldots, \theta_1) = \underline{\mu_1}$$
$$\mu_i(\theta_1, \theta_2, \ldots, \theta_1) = \underline{\mu_i} = \int_{-\infty}^{\infty} x^r f(x, \theta)dx; r = \underline{1, l}$$
$$\mu_k(\theta_1, \theta_2, \ldots, \theta_1) = \underline{\mu_k}$$

Оценки, полученные по методу моментов не оказываются эффективными. Альтернативой, дающей эффективные оценки, является метод правдоподобия. Постановка задачи совпадает. В основе этого метода лежит следующее дерево. Задана плотность: $f(x, \theta)$
Есть реализация $L(x_1, x_2, \ldots, , x_n, \theta\ ) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta)$ являющаяся статистикой выборки. Произведение в правой части называется **совместной плотностью распределения**, (см график [24]) Одна плотность характеризует генеральную выборку, а совместная плотность — конкретно нашу реализацию.

## Метод максимального правдоподобия

Мы получаем некоторую функцию, как либо убывающую / возрастающую, нам нужно чтобы эта функция была максимизирована/минимизирована, может быть несколько экстремумов.

$$L(x_1, x_2, \ldots, , x_n, \theta\hat{}) = maxL(x_1, x_2, \ldots, , x_n, \theta\ )$$

Функция L называется **функцией правдоподобия** и оценка $\theta\hat{}$ называется **оценкой максимального правдоподобия**.
[на всякий см 25]

$$\frac{dL(x_1, x_2, \ldots, , x_n, \theta\ )}{d\theta} = 0 \Rightarrow \theta\hat{}$$

Если $\theta$ многомерный параметр (вектор, $\theta\hat{} = (\theta_1, \theta_2, \ldots, \theta_l)$, тогда мы берем частную производную.

$$\frac{dL(x_1, x_2, \ldots, , x_n, \theta\ \hat{})}{d\theta_1} = 0$$
$$\frac{dL(x_1, x_2, \ldots, , x_n, \theta\ \hat{})}{d\theta_2} = 0$$
$$\frac{dL(x_1, x_2, \ldots, , x_n, \theta\ \hat{})}{d\theta_l} = 0$$

Распределение Вейбла:

$$F(x) = 1 - \alpha e^{\beta x^{\gamma} + \delta}$$

Метод даёт эффективные оценки, но оказывается вычислительно сложным.

# Задачи

## Схема Бернулли

$$P_m(A) = {c_n}^m p^m q^{n-m}$$
$${c_n}^m = \frac{n!}{m!\,(n-m)!}$$

Задача:
Радиолокационная станция. Вероятность исчезновения на радаре любого из 4-х объектов — 0,1.
Определить вероятность исчезновения:

- всех 4-х объектов одновременно (событие А)
- хотя бы одного объекта (событие Б)

Всех 4-х объектов:
$P(A) = P_4(1) + P_4(2) + P_4(3) + P_4(4)$— вероятность, что исчезнет
$P_4(1) = {C_4}^3 p^1 q^3$— вероятность, что исчезнет с радаров точно 1 объект
$P_4(4) = {C_4}^4 p^4 q^0$— вероятность, что исчезнет с радаров точно 4 объекта

$$P_4(0) = {C_4}^0 p^0 q^4$$
$$P(\text{Б}) = 1 - P_4(0)$$

## Метод моментов

$\xi_1 = (0, 0, 15, 23, 27, 33, 34, 40, 44, 49)$ — работа лампы в форсированном режиме
$\xi_2 = (12, 18, 34, 48, 48, 52, 54, 59, 60, 63)$ — работа лампы в нормальном режиме

$$\xi_1 = \varphi(\xi_2);\ \xi_1 = k\xi_2$$
$$\underline{\xi_1} = \frac{1}{n}\sum_{i=1}^{n} x_i = 26{,}5;\ \underline{\xi_2} = \frac{1}{n}\sum_{i=1}^{n} y_i = 44{,}3$$

$\underline{\xi_1} = \frac{k}{n}\sum_{i=1}^{n} y_i;\ \underline{\xi_2} = \frac{1}{n}\sum_{i=1}^{n} y_i \Rightarrow \underline{\xi_1} = k\underline{\xi_2};\ \hat{k} = \frac{\underline{\xi_1}}{\underline{\xi_2}} = \frac{26{,}5}{44{,}3} = 0{,}6$— оценочное k

$${\xi_1}^* = \hat{k}\xi_1 = (7, 11, 20, 25, 29, 31, 32, 35, 36, 38)$$
$$\underline{{\xi_1}^*} = 26{,}4 \sim 26{,}5 = \underline{\xi_1}$$

Но зависимость всё равно нелинейная. Разброс $\xi_1$составляет 49, а разброс $\xi_2$=51.
Выборки могут быть разного размера. То есть в $\xi_1$могло бы быть 8 элементов.

### Суббота (2013-10-12)

Пусть $(x_1, x_2, \ldots, x_n)$— независимая случайная повторная выборка объема n из генеральной совокупности.
Задана функция распределения $F(x, \theta)$ и плотность $f(x, \theta)$. $\theta$ — неизвестный параметр (можно интерпретировать его как детерминированную случайную величину или вектор).
Требуется построить интервал [ [teta минус снизу], [teta минус сверху] ], верхние границы которого являются элементы выборки.

$\theta\_ = f_i^{1}(x_1, x_2, \ldots, x_n)$ [teta минус снизу] = [fi 1] (x1, x2, …, xn)

Интервал $[\theta\_;\ \underline{\theta}]$ называется **доверительным интервалом** для параметра $\theta$ с коэффициентом доверия $\gamma$, если $\forall\theta: P(\theta\_ \le \theta \le \underline{\theta}) = \gamma$
$\gamma$ это параметр близкий к 1, больше 0.9 (0.9, 0.95….)

## Методы построения доверительных интервалов

Одним из часто используемых для построения интервалов методов является метод, основанный на предварительном построении некоторой центральной статистики. Центральной статистикой называется любая функции, зависящая от элементов выборки и параметра распределения: $T = f_i(x_1, x_2, \ldots, x_n, \theta)$

Предположим, что функция (центральная статистика) является монотонной функцией параметра $\theta$, например, монотонно убывающей.

Далее график [23]
**Квантилем уровня K** функции распределения$F(\alpha, \theta)$называется величина $\alpha = \alpha(K)$, определеяемая из условия $P(x \le \alpha) = F(\alpha) = K$.

**Постановка задачи**:
Определить доверительный интервал для математического ожидания случайной величины, распределённый нормально.

Зададимся двумя малыми числами $\epsilon_1, \epsilon_2$и определим величины $\alpha_1, \alpha_2$из условий:
Ищем: $\alpha_2 = \theta\_ \le \underline{\theta} = \alpha_1$

$$1 - \epsilon_2 = T(x_1, x_2, \ldots, x_n, \theta\_)$$
$$\epsilon_1 = T(x_1, x_2, \ldots, x_n, \underline{\theta})$$
$$P(x \le \alpha_1) = F(\alpha_1) = \epsilon_1 \Rightarrow P(x > \alpha_1) = 1 - F(\alpha_1) = 1 - \epsilon_1$$
$$P(x \le \alpha_2) = F(\alpha_2) = \epsilon_2 \Rightarrow P(x > \alpha_2) = 1 - F(\alpha_2) = 1 - \epsilon_2$$
$$\alpha_1 = \alpha(\epsilon_1)$$
$$\alpha_2 = \alpha(1 - \epsilon_2)$$
$$P(\alpha_2 < x \le \alpha_1) = P(x > \alpha_2) - P(x \le \alpha_1) = 1 - \epsilon_2 - \_\epsilon_1 = \gamma$$

В качестве примера можно привести оценку параметров нормального распределения

При известной дисперсии для доверительного оценивания матеожидания в качестве исходной центральной статистики берётся $T = \left(\frac{\underline{x} - \mu}{\sigma}\right)\sqrt{n}$ — среднее, а $\mu$— параметр.
Её специально выбирали (отдельная сложная задача). Статистика выбрана из соображений соответствия её стандартному нормальному распределению с параметрами (0, 1): 0 - теоретическое матожидания статистики T, 1 — дисперсия или среднеквадратичное отклонение.

Не существует формализованного подхода для выбора статистики.

Пример:

$$\left(\frac{\underline{x}-\underline{\mu_-}}{\sigma}\right)\sqrt{n}=\alpha(K_1)=\alpha(1-\epsilon)$$

$$\left(\frac{\underline{x}-\underline{\mu}}{\sigma}\right)\sqrt{n}=\alpha(K_2)=\alpha(\epsilon)$$

$$\epsilon_1=\epsilon_2=\epsilon=\frac{1-\gamma}{2}$$

$(\underline{\mu_-};\underline{\mu})$— интервал

$$T(x,\mu)=\alpha(K,\mu)=\alpha(K)$$

Ответ:

$$\underline{\mu_-}=\underline{x}-\frac{\alpha(1-\epsilon)\sigma}{\sqrt{n}}$$

$$\underline{\mu}=\underline{x}-\frac{\alpha(\epsilon)\sigma}{\sqrt{n}}$$

Статистика не заменяет функцию распределения, мы берём её лишь удобную для подсчёта.

Отрицательные качества:

1. 
2. Мы испытываем проблемы с нормальным распределением (оставляется $\sqrt{n}$)

Положительные качества:

1. Для типовых заданий имеет готовое решение
2. Не нужно иметь несколько статистик. Мы всего лишь смотрим доверительный интервал для одной статистики. Хороший практический результат.

## Пример решения методом моментов

**Дано:**

$\xi_1$— случайная величина (показатель 1)

$\xi_2$— случайная величина (показатель 2)

**Найти:**

$\hat{a},\hat{b}$ используя метод моментов

$$\xi_1=a{\xi_2}^b\Rightarrow ln\xi_1=lna+bln\xi_2$$

**Решение:**

Берём выборочное среднее. Переобозначим переменные $\psi_1=A\psi_2+B$; $\frac{1}{n_1}\sum_{i=1}^{n_1}\quad\psi_{i1}=\frac{A}{n_2}\sum_{i=1}^{n_2}\quad\psi_{i2}+B$или $\frac{1}{n_1}\sum_{i=1}^{n_1}\quad ln\xi_{i1}=\frac{b}{n_2}\sum_{i=1}^{n_2}\quad ln\xi_{i1}+lna$ Избавляемся от константы B, так как матожидание равно B, а дисперсия равна 0

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\quad\left(\psi_{i_1}-\underline{\psi_1}\right)^2=\frac{b^2}{n_2}\sum_{i=1}^{n_2}\quad\left(\psi_{i_2}-\underline{\psi_2}\right)^2$$

$$M_1=bM_2+lna$$

$$D_1{}^2 = bD_2{}^2$$

Из 2 вышестоящих уравнений: $\hat{b} = \sqrt{\frac{D_1{}^2}{D_2{}^2}}$; $\hat{a} = e^{(M_1 - bM_2)}$

Матожидание: $M_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \quad ln\xi_{i1}$;$M_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \quad ln\xi_{i2}$

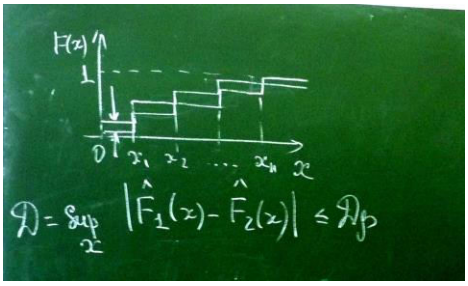
Дисперсия: $D_1{}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1} \quad (ln\xi_{i1} - M_1)^2$;$D_2{}^2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \quad (ln\xi_{i2} - M_2)^2$

$\xi_1{}^* = \hat{a}\xi_2{}^{\hat{b}}$— функциональная зависимость между 2-я выборками

$\xi_1{}^*$ — расчётная выборка

$\xi_2$— контрольная выборка. Но так не правильно, потому что подгонять нужно по одной выборке, а контролировать по другой. Решение - выборку $\xi_1$разделяем на 2 (чётные и нечётные значения)

Строим функции распределения с одинаковым шагом




для расчётной $\xi_1{}^*$ и контрольной $\xi_1{}^{(2)}$ выборок.

Смотрим разницу:

$D = sup_x|\widehat{F_1}(x) - \widehat{F_2}(x)| \leq D_p$, где $D_p$— соответствует определённому уровню значимости.

f

# Основные понятия теории проверки статистических гипотез

**Статистической гипотезой H** называют любое утверждение относительно функции распределения F(x) случайной величины x, касающееся типа функции распределения, значения её параметров и т. п.

Гипотеза H проверятся путём сопоставления выдвинутых предположений с результатами экспериментов.

В статистике результаты эксперимента представляют собой выборку объёма n из некоторой генеральной совокупности, независимой выборочной функцией распределения приближается теоретической функции распределения F(x).

Гипотезу называют **простой**, если её условием удовлетворяет единственная функция распределения F(x). В противном случае гипотеза называется **сложной**.

Гипотезу, справедливость которой определяется (устанавливается) в ходе эксперимента, называют **нулевой** или **основной** гипотезой $H_0$. В зависимости от того, какие отклонения возможны от гипотезы $H_0$ формулируют **альтернативные** (**конкурирующая**) гипотезы $H_1$.

Статистические гипотезы проверяют посредством статистических критериев.

**Статистический критерий** — совокупность правил, позволяющих по полученной выборке принять гипотезу $H_0$ или отвергнуть её в пользу гипотезы $H_1$.

Имеется общий подход построения статистического критерия:

1. Выбирается статистика исследователем $S = S(x_1, x_2, \dots, x_n)$, $S \subset \Omega$
2. Множество $\Omega$ всех возможных значения статистики разбивают на 2 подмножества: $T_0$ — множество принятия решений гипотезы $H_0$ и $T_{\text{кр}}$ — критическое множество (множество больших, чем альтернативные гипотезы $H_1$)
3. Если полученное значение статистики попадает в множество $T_0$ ($S^{(1)} = S({x_1}^{(1)}, {x_2}^{(1)}, \dots, {x_n}^{(1)}) \in T_0$), то гипотезу $H_0$принимают. Если значение статистики принадлежит $T_{\text{кр}}$ ($S^{(2)} = S({x_1}^{(2)}, {x_2}^{(2)}, \dots, {x_n}^{(2)} \in T_{\text{кр}})$), то гипотеза $H_0$отвергается в пользу $H_1$

В силу того, что $S(x_1, x_2, \dots, x_n)$— случайная величина, принадлежность этой величины критическому множеству может произойти как при справедливости $H_1$, так и при справедливости $H_0$. В связи с этим возможны ошибки 2 типов:

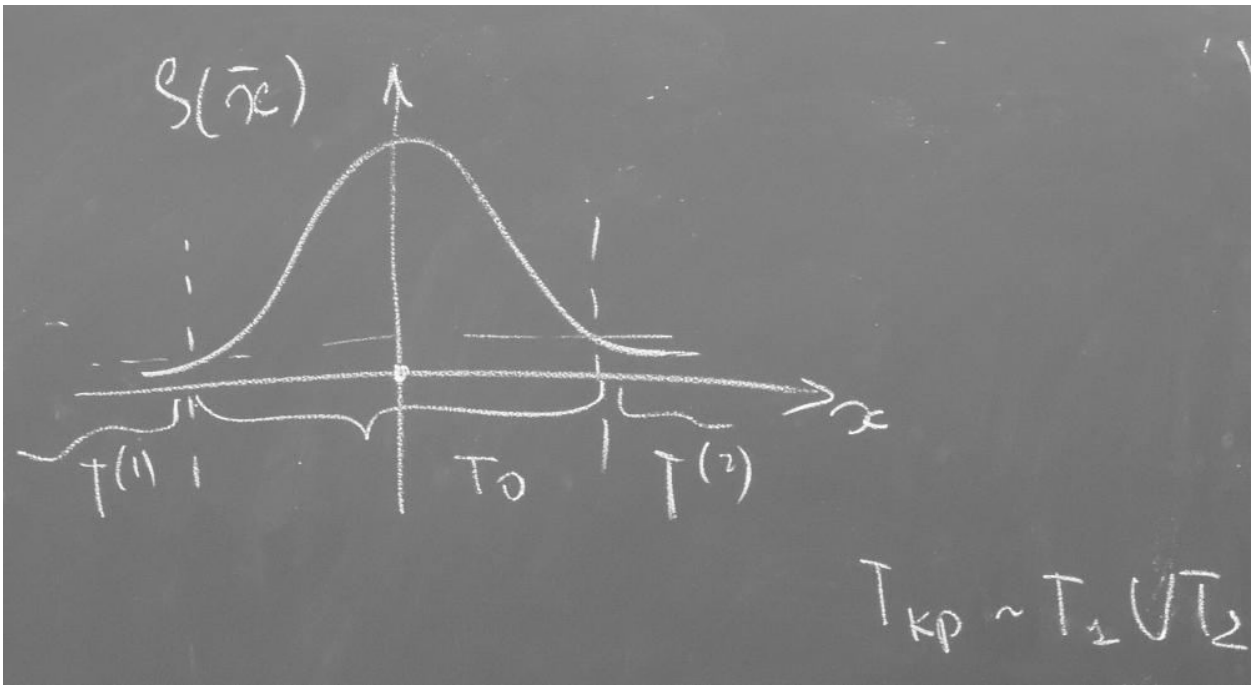
**Ошибка 1 рода**: $H_0$верно, но принимается $H_1 (S(x_1, x_2, \dots, x_n) \in T_{\text{кр}})$при условии
**Ошибка 2 рода**: $H_1$верно, но принимается $H_0 (S(x_1, x_2, \dots, x_n) \in T_0)$

**Вероятность ошибки 1-го рода** обозначается $\alpha = P(S(x_1, x_2, \dots, x_n) \in T_{\text{кр}} \mid H_0)$
**Вероятность ошибки 2-го рода** обозначается $\beta = P(S(x_1, x_2, \dots, x_n) \in T_0 \mid H_1)$

Вводится понятие **мощности критерия** $\gamma = 1 - \beta$ — вероятность правильно отвергнуть $H_0$.

Так как нельзя одновременно уменьшать значение $\alpha$ и $\beta$на практике поступают следующим образом: фиксируют $\alpha$, максимизируя при этом $\gamma$. Нельзя одновременно уменьшать $\alpha$ и $\beta$, так как $\alpha + \beta = const$

Двусторонний критерий: (разбился на 2 подмножества) $T_{\text{кр}} \sim T_1 \cup T_2$




В противном случае, критерий односторонний.

Выбор одностороннего или  двустороннего критерия определяется видом альтернативной гипотезы.

Множество, отделяющее основную и альтернативную гипотезы, называется **зоной индифферентности**.

фото

Если заранее известен закон распределения случайной величины x, статистические выводы могут быть точнее, но требуется показать, на сколько результаты эксперимента соответствуют предположению о законе распределения. С этой целью используют критерии согласия.

## Критерии согласия

**Критериями согласия** называют критерии, в которых гипотеза определяет закон распределения полностью, либо с точностью до небольшого числа параметров.

Пусть имеем выборку $x_1, \ldots, x_n$. Теоретическую функцию (реальную, но неизвестную) распределения обозначим $G(x)$, а гипотетические (по гипотезе) $F(x)$. При этом эмпирическая (выборочная) функция распределения — $\widehat{F_n}(x)$. Тогда гипотеза $H$ о том, что истинное распределение $G(x)$есть $F(x)$ записывается следующим образом $H\colon G(\cdot) = F(\cdot)$.Пишем с точками потому, что они могут совпадать на некотором интервале(ах) области определения, а не на всём области определения — потому что берём значения лишь ограниченные в некоторой области.

Если гипотеза H верна, то $\widehat{F_n}(x) \to F(x)$ при $n \to \infty$
Если гипотеза H не верна, то $\widehat{F_n}(x) \to G(x)$при $n \to \infty$, но мы эту $G(x)$не знаем.

## Критерий Колмогорова-Смирнова

Наиболее распространённой статистикой при проверке статистической гипотезы H, является **статистика Колмогорова**: $D_n = sup|F(x) - \widehat{F_n}(x)|, -\infty < x < \infty$.

При малых n для гипотезы H составлены (процентные) таблицы, из которых выбирается предельное значение $D_\beta$, соответствующих заданному уровню значимости $\beta$(ошибки 2-го рода). Критерии вида $sup|F(x) - \widehat{F_n}(x)| < D_\beta, -\infty < x < \infty$ называются критериями Колмогорова-Смирнова.

Используют эмпирическую формулу: $D_\beta = \sqrt{-\frac{1}{2} ln(1-\beta)}$— не очень точна.

Где взять $D_\beta$? $D_n = sup\left(\frac{k}{n} - F(x_k);\ F(x_k) + \frac{k-1}{n}\right), 1 < k < n$ Разбиваем на маленькие отрезки и берём sup.

## Критерий $\omega^2$

$$\omega_n{}^2 = \int_{-\infty}^{\infty} (F(x) - \widehat{F_n}(x))^2 dF(x)$$

Для вычисления по реальной выборке используют формулу, полученной тоже эмперически:

$$n\omega_n{}^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F(x_i) - \frac{2i-1}{2n}\right)^2$$

Если эта масштабированная величина не превышает какого-то значения, то признаём соответствие гипотезе $H_0$.
По аналогии для критерия $\omega^2$ составлены процентные таблицы для правой части (ограничения, критерия) для выборок небольшого размера.

## Для сложных гипотез

Сложная гипотеза — то есть проверяется соответствие закону распределения по неизвестным параметрам, то есть имеем семейство
Более сложной задачей является проверка сложной гипотезы. Например, подтверждение наличия некоторого распределения при векторе неизвестных параметров. В этом случае работают с оценкой вектора неизвестных параметров, например, полученных с помощью методов максимального правдоподобия.
То есть если мы для каждого эксперимента имеем аппроксимируемую кривую, но хотим считать как одно. Для этого их “объединяем” методом максимального правдоподобия.
Фиксируем параметры и избавляемся от семейства, оставляя только простую гипотезу.
То есть $F(x, \theta) = F(x)$

Пусть $\hat{\theta}$— оценка вектора неизвестных параметров, тогда модифицированная статистика критерия Колмогорова-Смирнова $\widehat{D_n} = sup_x|F(x,\theta) - F(x,\hat{\theta})| \leq D_\beta$. А модифицированная статистика критерия $\omega^2$выглядит соответственно: $\widehat{\omega_n} = \int_{-\infty}^{\infty} \quad |F(x,\theta) - F(x,\hat{\theta})|dF(x,\hat{\theta})$.

## Ранговые критерии

Критерий согласия Пирсона в ряде случаев называется критерием $\chi^2$. Достаточно часто употребляемый критерий для проверки гипотезы о принадлежности наблюдаемой выборки $(x_1, x_2, \ldots, x_n)$ некоторой генеральной совокупности, распределённой по закону $F(x,\theta)$. В случае простой гипотезы выборочная функция распределения соответствует эмпирической $F_n(x) = F(x,\theta)$, где $F_n(x)$— выборочная функция распределения.
Есть выборка, которая разбивается на группы по определённым критериям. Каждой группе назначается ранг. Оценивается именно ранг, а не сами выборки. Ранг вбирает в себя информацию по нескольким значениям.
Процедура проверки гипотез с использованием критериев типа $\chi^2$предусматривает группирование наблюдений. Область определения случайной величины $x$разбивается на k непересекающихся групп граничными точками $x_{(0)}, x_{(1)}, x_{(2)}, \ldots, x_{(k)}$. Вводится нижняя грань $x_{(0)} = x_1$— минимальное значение вариационного ряда и верхняя грань $x_{(k)} = x_n$ — максимальное значение вариационного ряда.
Вычисляется $n_i(\sum_i \quad n_i = n)$, попавших в каждый i-й интервал и вероятности попадания в интервал $P_i$, соответствующие теоретическому закону распределения $F(x,\theta)$
$P_i(\theta) = F(x_{(i)},\theta) - F(x_{(i-1)},\theta)$, при этом $\sum_i \quad P_i(\theta) = 1$

Пример:
Область определения (0;20) разбили на 4 равномерные части (цвета):

| x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 4 | 5 | 7 | 8 | 12 | 14 | 15 | 17 | 18 |

Эмпирическим аналогом вероятностей $P_i(\theta)$ является отношение $\frac{n_i}{n}$, то есть в основе статистик, используемых в критериях согласия типа $\chi^2$, лежит изменение отклонений этого отношения от $P_i(\theta)$.

### Критерий Пирсона

$$\chi^2{}_n = n\sum_{i=1}^{k} \quad \frac{\left(\frac{n_i}{n} - P_i(\theta)\right)^2}{P_i(\theta)}$$

*$n_i$- количество чисел, попавших в интервал. Например, в интервал x2-x1 попало 2 числа (см таблицу-пример выше).*
*n - размер выборки*
*k - число групп*

*Pi -*

Статистика критерия согласия $\chi^2$ при $n \to \infty$подчиняется распределению $\chi^2$ с $k-1$ степенями свободы, если верна 0-я гипотеза (выборочное распределение совпадает с теоретической при $n \to \infty$).
Распределение $\chi^2$с $k-1$ степенью свободы — это распределение суммы квадратов $k-1$ независимых стандартных ($\mu = 0, \sigma = 1$) нормально распределённых случайных величин.
$k-1$ пришло из теоретической механики

## Практическое применение критерия Пирсона

**Дано** сведения по защитам дипломных проектов в ВУЗе за определённый временной промежуток.

| | отл, % | хор, % | удов, % | общее кол-во выпускников |
|---|---|---|---|---|
| 2003 | 68 | 25 | 7 | 1485 |
| 2004 | 40 | 40 | 20 | 1412 |
| 2005 | 25 | 33 | 12 | 1388 |
| 2006 | 59 | 28 | 13 | 1435 |
| 2007 | 48 | 37 | 15 | 1422 |

Определить, существует ли зависимость между количеством выпускников и распределением оценок при защите дипломных проектов (признак квалификации).

### Решение

Нужно сформулировать 0-гипотезу и альтернативную. В нашем случае считаем, что оценки не зависят от количества выпускников. Значит 0-гипотеза — что не зависит, альтернативная — зависимость присутствует.

Решим более простую задачу.

$$\xi_1 = (40, 48, 44, 55, 59, 68)$$

| x(0)=40 | x(1)=46 | x(2)=52 | x(3)=58 | x(4)=64 | x(5)=70 |
|---|---|---|---|---|---|

| x(0)-x(1) | x(1)-x(2) | x(2)-x(3) | x(3)-x(4) | x(4)-x(5) |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| ⅕ | ⅕ | ⅕ | ⅕ | ⅕ |

$$\xi_2 = (1388, 1412, 1422, 1435, 1485)$$

| x(0)=1385 | x(1)=1405 | x(2)=1425 | x(3)=1445 | x(4)=1465 | x(5)=1485 |
|---|---|---|---|---|---|

| x(0)-x(1) | x(1)-x(2) | x(2)-x(3) | x(3)-x(4) | x(4)-x(5) |
|---|---|---|---|---|
| 1 | 2 | 1 | 0 | 1 |
| ⅕ | ⅖ | ⅕ | 0 | ⅕ |

В качестве $P_i$ возьмём первый ряд. В итоге, $\chi^2 = 2$
Дальше строится распределение [0; 1] и проверяется

## Лабораторная

Есть 2 выборки. Если после ранжирования присутствует вероятность в каждом ряде равной 0, то просто уменьшаем количество групп.
Сравнить по Пирсону и по Спирману.

## Критерий Спирмена

Альтернативы критерию Пирсона является критерий, использующий статистику Спирмэна.
Дано:

1. Данные по защите дипломных проектов группы N вуза U.
2. Данные по оценкам абитуриентов, полученных по результатам ЕГЭ

Проверить гипотезу о независимости оценок за диплом и оценку за ЕГЭ. 0-гипотеза — независима, альтернативная — зависимы.

Статистика:
$\rho_s = 1 - \frac{1}{n(n^2-1)} \sum_{i=1}^{n} \quad (R_i - S_i)^2$, где $R_i, S_i$— ранги двух последовательностей, n — одинаковый объём выборок
Здесь ранг - порядковый номер интервала.

Для этой задачи получилось, что $\rho_s = 0.95$
Нормируем (восстанавливаем): $\sqrt{n-1}\rho_s \sim N(0,1)$, где $N$— нормальный закон распределения. Получаем значение 4.15
Из таблицы критических значений выбирается строка, соответствующая объёму выборки N=20 и столбец, соответствующий уровню значимости 0,05 . Величина, находящаяся на пересечении соответствующей строки и столбца, определяет границу критической области.
4 ранга для дипломного проектирования: 2,3,4,5
9 рангов по вступительным испытания

(таблица на след. странице для удобства чтения).

| | Дипломное проектирование | Ранг оценки Ri | Вступительные Испытания | Ранг ВИ Si | $(R_i - S_i$ |
|---|---|---|---|---|---|
| | 5 | 4 | 4,25 | 5 | 1 |
| | 5 | 4 | 4,50 | 6 | 4 |
| | 4 | 3 | 3,25 | 2 | 1 |
| | 5 | | 4,25 | 5 | 1 |
| | 4 | | 4,50 | 6 | 9 |
| | 4 | | 4,00 | 4 | 1 |
| | 3 | | 3,00 | 1 | 1 |
| | 4 | | 3,50 | 3 | 0 |
| | 4 | | 3,50 | 4 | 0 |
| | 2 недопуск | | 3,25 | 2 | 1 |
| | 4 | | 4,00 | 4 | 1 |
| | 4 | | 3,50 | 3 | 0 |
| | 5 | | 4,50 | 6 | 4 |
| | 5 | | 4,75 | 7 | 9 |
| | 3 | | 3,0 | 1 | 1 |
| | 5 | | 4,25 | 5 | 1 |
| | 5 | | 4,75 | 7 | 9 |
| | 5 | | 5,00 | 8 | 16 |
| | 4 | | 4,00 | 4 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| | 3 | | 3,25 | 2 | 0 |

# Случайные функции

Скалярная или векторная функция от переменной $t \in T$, принимающая на области определения T, являющаяся случайными величинами с некоторой функцией , распределения, называется случайной функцией $\xi(t)$
Вектор $\underline{\xi}(t) = \xi_1(t, \alpha_1);\ \xi_2(t, \alpha_2), \ldots, \xi_n(t, \alpha_n)$

Если в качестве аргументы используется переменная времени, то говорят о **случайном процессе**.

**Марковский процесс** —- случайный процесс, развитие которого в данный момент времени t не зависит от состояния системы в предшествующие моменты времени, если значение $\xi\ \ (t)$в предыдущий момент времени фиксировано.
Другая трактовка по Венцель: “Будущее” процесса не зависит от “прошлого”, при известном “настоящем”. Или: “Будущее” процесса зависит от “прошлого” через “настоящее”.
Различают марковскую цепь:

- с непрерывным временем — время непрерывно, пространство состояний — дискретно.
- и с дискретным временем. Время дискретно и пространство состояний тоже дискретно.

**Марковский процесс** — понятие используемое, когда и время, и пространство состояний непрерывны.

**Цепь Маркова** — последовательность случайных событий с конечным или счётным числом исходов, характеризующееся свойством независимости будущего от прошлого при фиксированном настоящем.

**Однородная марковская цепь** называется такая цепь Маркова, вероятность перехода которой из одного состояния в другое не зависит от времени. Если зависит, то цепь называется **неоднородной**.

При постановке задачи определяются графически или таблично переходы из одного состояния системы в другое (задаются наборы состояний) и вероятности переходов.

**Лабораторная 4**

**Дано**: Стрельбы $S = \{S_0, S_1, S_2\}$
$S_0$ — неповреждённая мишень
$S_1$ — неповреждённая мишень
$S_2$ — разрушенная мишень

| | $S_0$ | $S_1$ | $S_2$ | |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| $S_0$ | $P_{00}$ | $P_{01}$ | $P_{02}$ | $\sum_i P_{0i} = 1$ |
| $S_1$ | 0 | $P_{11}$ | $P_{12}$ | $\sum_i P_{1i} = 1$ |
| $S_2$ | 0 | 0 | 1 | $\sum_i P_{21} = 1$ |

| | $S_0$ | $S_1$ | $S_2$ |
|---|---|---|---|
| $S_0$ | 0,2 | 0,7 | 0,1 |
| $S_1$ | 0 | 0,6 | 0,4 |
| $S_2$ | 0 | 0 | 1 |

Выше — матричное представление графа переходов. Можно же визуализировать этот граф.

**Найти**: Среднее количество боеприпасов, необходимых для поражения (уничтожения) цели.

Решение:

1. Запускаем генератор случайных числе, равномерно распределённый на отрезке (0, 1) и получаем последовательность: 0,24; 0,63; 0,86; 0,44; ...

$$S_0 \to S_1 \to S_2$$

2. 0,15; 0,28; 0,46; 0,68

$$S_0 \to S_0 \to S_1 \to S_{\ 1} \to S_2$$

3. 0,13; 0,23; 0,36; 0,54; 0,70

$$S_0 \to S_0 \to S_1 \to S_1 \to S_1 \to S_2$$

4. 0,21; 0,53; 0,71; 0,94

$$S_0 \to S_1 \to S_1 \to S_2$$

5. 0,92

$$S_0 \to S_2$$

6. 0.76; 0.11; 0.28; 0.94; 0.53; 0.88;
7.

Количество выстрелов: 2, 4, 5, 3, 1

Результатом первой итерации явилось среднее количество боеприпасов, необходимы для поражения цели, равное 3. Необходимо уточнить количество экспериментов, требуемых для получения эффективной оценки параметров.

График: Ось y - среднее количество выстрелов из x последовательностей

Критерий: 3 последний значения должны попасть в диапазон. То есть строим график (количество последовательностей - среднее количество выстрелов). Потом считаем среднее между последними 3-мя точками (n-2, n-1, n выстрелов) и сравниваем со средним между предпоследними 3-мя точками (n-3, n-2, n-1 выстрелов).

Лабораторная: Выбрать предметную область, где можно оценить вероятность переходов.

# Моделирование случайных величин и векторов с заданным законом распределения

Как правило, рассматривается следующая классификация ГСЧ (генераторов случайных чисел):

- алгебраические
- табличные
- физические

**Физический ГСЧ** (любой случайный процесс в природе).

Примером физических генераторов являются генерация k-разрядных случайных чисел подбрасыванием монеты, игральной кости, вращением волчка. Регистрация колебаний любой природы (переменное напряжение, звуковое сообщение). То есть величины при дискретизации колебаний.

**Генераторы табличные**

К табличным ГСЧ относят таблицы специального вида, представляющих собой набор независимых (это строго проверяется) друг от друга случайных чисел, интерпретируемых как цифры от 0 до 9. Комбинация таких цифр будет представлять собой сгенерированное случайное число.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | 2 | 9 | 2 | 0 | 4 | 2 | 6 |
| 9 | 0 | 7 | 3 | 1 | 9 | 0 | 3 |
| 5 | 9 | 1 | 6 | 6 | 5 | 5 | 7 |

В клетках — случайные числа.

Выбираем точность и последовательно получаем: 0,929; 0,204; 0,269; 0,073; 0,190; 0,359; 0,166; 0,557.

## Алгебраические

Самый широкий класс — это алгебраические ГСЧ. Для реализации такого генератора требуется

1. выбор стартового значения
2. произвольный алгоритм генерации псевдослучайного числа
3. правило зацикливания алгоритма

Примеры:

Берём середину (4 цифры) и возводим квадрат. Повторяем:

1. 12**0863**17 = 0,0863
2. 00**7447**69 = 0,7447
3. 55**4578**09 = 0,4578
4. 20**9580**84 = 0,9580
5. 91**7764**00 = 0,7764
6. 60**2796**96 = 0,2796
7. 07**8176**16 = 0,8176
8. 66**8469**76 = 0,8469

Перемножаем 2 числа, делаем сдвиг по разрядам и разделяем:

1. R1=1023 R2=6324: R1*R2 = 06***4694***52. Сдвигаем влево на 1 разряд и вправо. Получим числа: 64**6945**20, 20**6469**45. Выбрали случайное число: 0, 4694
2. 6945*6469=44***9272***05. После сдвига: 49**2720**54, 54**4927**20. Выбрали случайное число: 0,9272
3. 2720*4927=13***4014***40. После сдвига: 34**0144**01, 01**3401**44. Выбрали случайное число: 0,4014
4. 0144*3401=00***4897***44. После сдвига: 04**8974**40, 40**0489**74. Выбрали случайное число: 0,4897
5. 8974*0489=04***3882***86. После сдвига: 43**8828**60, 60**4388**28. Выбрали случайное число: 0,3882

Чем лучше этот метод по отношению с предыдущим? Тем, что нет появления 0 и прямого зацикливания.

Известно достаточно много модификаций этого метода с разными периодами зацикливания, не лишённые проблемы зацикливания в принципе. Кроме того, появление в качестве стартового значения числа 0000 делает алгоритм неработоспособным, в связи с чем в современных инструментальных средствах применяют так называемый **конгруентный метод**.

Конгруентный метод предполагает построение рекурсивной зависимости вида: $x_{i+1} = mod(ax_i + b, \mu)$. a и b выбираются эмпирически, а модуль $\mu = 2^N$. Обычно N связывают с разрядностью машины.

## Генерация случайных величин по заданному закону распределения

### Нормальный закон распределения

Формула Мюллера:

Пусть имеются значения $x_1, x_2$ равномерно распределённых на отрезке [0; 1]. Тогда y — случайная величина, распределённая нормально, моделируется по формуле: $y = \mu + \sigma\sqrt{-2lnx_1}sin2\pi x_2$

На основе центральной предельной теоремы.

Генерируют 6 значений ($n = 6$), распределённых равномерно на [0; 1]. В этом случае $\mu_1 = \frac{n}{12}$; $\sigma = \sqrt{\frac{n}{12}}$

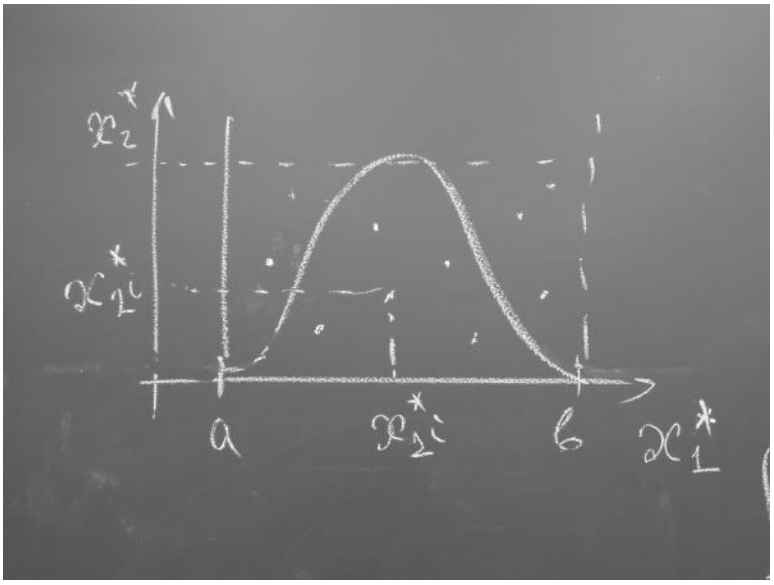
??????

---

Метод Неймана генерации любой случайной величины по любому случайному закону

$x_1, x_2$ — равномерно распределено на [0; 1]

$y \in [a;\ b]$ случайная величина с заданным законом распределения




$$x_1{}^* = a + bx_1$$

$$x_2 *= f_{max}x_2$$

$$f_{max} = \max_y f(y)$$

Оставляем точки случайных значений, попадающие под график нормального распределения. При этом точки, выше графика убираем.

$x_{2i}{}^{*} \leq f(x_{1i}{}^{*})$, то оставляем это значение, иначе отбрасываем.

$m_1$ - мат. ожидание случайной величины, распредленной равномерно на отрезке [0, 1]

$m_v$= n * $m_1$= n/2 (теоретическое значение, математическое ожидание нормально - распределенной случайно величины, полученной суммированием n случайных величин, равномерно распредленных на отрезке [0,1])

! Повторить “**Закон больших чисел**” и “**Центральная предельная теорема**”

$M_v = n \cdot m_i = \frac{n}{2}$

Ряды:

$V_1 = 5,5$ 0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1,0 $M_1 = 0,5$

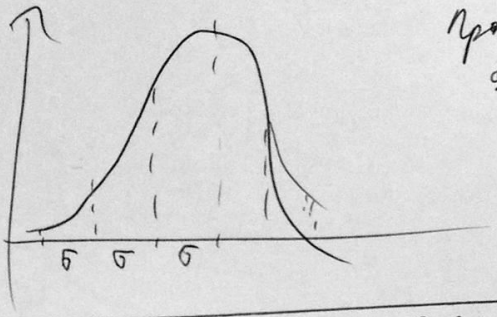
$V_2 = 3,0$ 0; 0,2; 0,4; 0,6; 0,8; 1,0 $M_2 = 0,5$

3. Центральная предельная теорема
Теория больших чисел

будет в основном знания

Правило 3σ

95% входит в 6σ (по 3 с каждой стороны)




$(x_1, x_2, \dots, x_n)$ $f(x)$ $F(x) = \int_{-\infty}^{+\infty} f(x)\,dx = y$

Метод обратной функции распределения (преобразования)

$f(x) = \lambda e^{-\lambda x} \Rightarrow F(x) = 1 - e^{-\lambda x}$, $x \geq 0$

$1 - e^{-\lambda x} = y$

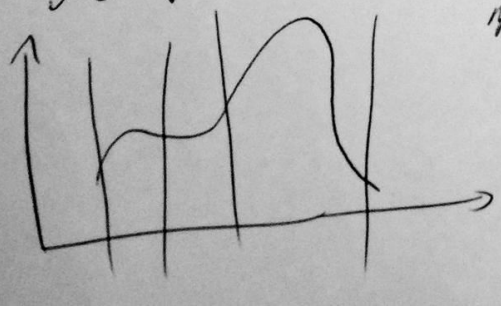
$\ln(1 - e^{-\lambda x}) = \ln y$

$\ln(e^{-\lambda x}) = \ln y$ – упрощение формулы возможно, т.к. случайные величины z и случ. велич. 1-z имеют один и тот же равномерный закон распределения. (просто разные диапазоны)

$-\lambda x = \ln y \Rightarrow x = -\frac{1}{\lambda} \ln y$ (из равномерно распредел. у получаем х, распредел. по показат. закону)

Метод суперпозиции

приближаем на участках



Альтернативным методу Неймана, является **метод обратного преобразования** или **метод обратной функции распределения**. Идея его заключается в том, что для выборки ($x_1$,$x_2$,...,$x_n$) определена плотность распределения f(x), для которой вычисляется функция распределения

F(x) = integral(f(x)dx) = yi (некоторая случ. вел. распр. равномерно в некотором диапазоне);

Тогда случайная величина имеет заданный F(x) закон распределения, определяемый аналитически, как ф-ция y(x), при условии, что случайная величина y равномерно распределена в интервале.

Пример (показательно распределение):

f(x) = lambda * e ^ (-lambda * x), x >= 0, -> F(x) = 1 - e ^ (-lambda * x), x >= 0;

1 - e ^ (-lambda *x ) = y  =>

ln(1 - e ^ (-lambda * x) )= ln(y)

ln(1 - e ^(-lambda * x)) = ln(1 - y) => -lambda*x = ln(y) => x = (-1/lambda) * ln(y)

Такое упрощение формулы возможно, так - как случайные величины z  и случайная величина 1 - z имею один и тот же равномерный закон распределения.

Т.е. имея набор случайных величин y, равномерно распределнных на интервале [0,1] получаем выборку случайных величин x, распределённых по экспоненциальному закону с параметром распределения lambda.

Возможно получить аналитическую зависимости для случайных величин, распределенных по закану arcsin и закону Коши, используя свойство симметрий тригонометрических функций

Кроме того, метод обратной функции используют для закона Релея. Остальные функции распределения дают сложный вид зависимости и не используются практически.

Достаточно универсальным (для любого вида распределения) и не имеющим ограничений (усечения функции плотности распределения) является метод суперпозиций или метод суммирования.

Основная идея метода суперпозиции заключается в приближении функции плотности распределения простыми (в плане описания) законами. Т.е. случайная величина x в разных диапазонах будет моделироваться разными алгоритнами.

Леция 23.11

# Метод многомерного статистического анализа

Обзоры методов, поскольку методов достаточно много.

Каждый объект в выборке может содержать наблюдения в более чем над одной случайной переменной (в этом случае все переменные считаются случайными и изучаются взаимодействия между зависимой переменной и совокупностью независимых переменных). Все методы многомерного статистического анализа, за исключением регрессии анализирует все переменные одновременно как случайный вектор с многомерным распределением.

Многомерный анализ это анализ множественных результатов измерений свойств (характеристик) объектов случайной выборки различными методами подстановки спектра задач.

рис №1.

*Пример:*
*В отделение терапии поступили больные, делаем их полный осмотр. На основании показателей, понимаем, в каком состоянии пациент и принимаем решения.*
*Т.е. объект - пациент*
*свойства - кровь, давление, температура, итд.*
*Нужно классифицировать объекты по группам (среднее, тяжелое, реанимация). Исходя из классификации помочь больному.*

Будем классифицировать задачи анализа временных рядов, факторного анализа, а также задачи классификации и кластеризации.

Временной ряд это последовательность измерений значений переменной (процесса, когда его дискретизируем, например через равные промежтки времени снимаются с датчика данные) произведенные через определенные, чаще равные, значения параметра (времени). Если измеряемые значения являются многомерными, то и временной ряд является многомерным. Область статистики, анализирующая временные ряды, называется анализом временных рядов.

Задачи, которые решаются с помощью временных рядов.
На основе АВР (анализ временных рядов) решаются следующие классы задач:

1. Построение модели объекта или процесса, описываемого временным рядом.
2. Для исследования структуры временного ряда (выявление тренда, анализ среднего выборочного среднего уровня значения тренда, обнаружение периодических колебаний, выявление лагов, задержек).
3. Вытекает из пункта 2, уже построили модель. Задача прогнозирования будущего развития процесса, представленного временным рядом.

Совокупность перечисленных задач (и прочих) решается следующими методами:

1. Методы корреляционного анализа (позволяет выявить наиболее существенные периодические зависимости, лаги, периоды)
2. Методы спектрального анализа (сигнал раскладывается на составляющие)
3. Методы сглаживания и фильтрации, предназначенные для преобразования временных рядов, устранения выбросов.
4. Методы авто-регрессии и скользящего среднего.

## Факторный анализ

При исследовании сложных объектов и систем часто невозможно напрямую измерить величины, определяющие свойства этих объектов, а иногда неизвестны смысл и значение измеряемых величин. Зато доступны для измерения величины в той или иной степени зависящие от исследуемой величины.

> **Примеры:**
>
> Можем померить температуру стенок реактора, но не можем измерить внутри. Но мы можем построить функцию, которая была бы внутри. Но для внешней среды термоса это не информативно.
>
> Можем посмотреть спектр, но можем не понимать, что означает краснота в этом спектре.

Величины, подлежащие измерению, называют **факторами эксперимента**.

Для обнаружения влияющих на измеряемые переменные факторов применяются методы факторного анализа. То есть, когда влияние неизвестного фактора проявляется в нескольких измеряемых факторах (признаках), эти признаки могут обнаруживать тесную связь между собой (кореллированность), что позволяет уменьшить общее количество измеряемых исследователем величин первоначально выбираемых произвольно.

Наиболее распространённым вариантом факторного анализа является является одно/двух факторный анализ с известными заранее факторами. Выбор новых признаков, которые являются линейными комбинациями других, осуществляется на основе вспомогательных методов (метод на основе метода). В частности, на основе метода главных компонент.

> Принцип метода: Чем больше разброс по фактору, тем он информативнее. То есть выбираем фактор, имеющий наибольший разброс.

Как было сказано, факторный анализ с применением выбора главных компонент хорошо работает в случае известных факторов. Хотя часто выбранные факторы интерпретируются неоднозначно. Для интерпретации факторов проводится причинно-следственный анализ компонентов. Дальнейшим исследованием применяются только факторы, подлежащие интерпретации.

## Дискриминантный анализ

Пусть имеется совокупность объектов, разбитая на несколько групп (для каждого объекта можно сказать к какой группе он относится). Пусть для каждого объекта имеется совокупность количественных характеристик. Для вновь появившегося объекта на основе значений его признаков необходимо выбрать группу, к которой он относится.

Для решения этой задачи применяют методы дискриминантного анализа, которые позволяют строить функции измеряемых характеристик, значения которых объясняют разбиение отрезков на группы.

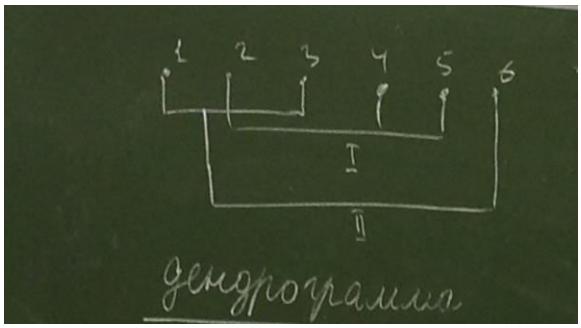
1. Желательно небольшое количество дискриминирующих признаков.
2. Особую роль играет линейный дискриминантный анализ, в котором классифицирующие признаки выбираются как линейные функции от первичных признаков (проще считать). Часто речь просто идёт о масштабировании.

Единственным ограничением метода является отсутствие предварительной информации об объектах, относящихся к конкретной группе. В этом случае используют методы кластерного анализа, позволяющие разбить изучаемую совокупность объектов на группы схожих объектов, называемых кластерами.

Почему иногда требуется не дискриминантный анализ, а кластерный. Кластерный нужен для тех случаев, когда мы не знаем группы (н интерпретируются группы). В случае, когда группы известны, то следует применят дискриминантный анализ.

## Кластерный анализ

Большинство кластерных методов являются аггломеративными (объединительными). Предполагается пошаговое объединение элементарных объектов с последующим объединением групп объектов.




Можем объединить кластеры, как на рисунке.

В этом случае строится дерево объединения кластеров, называемое **дендрограмма**.

Другой тип методов предполагает разбиение кластеров и называется **дивизивным** (разделительным).

На основание условия разделения базовый кластер разбивается на два кластера. На последующих шагах процедура повторяется, до достижения заданного числа кластеров.

У кластерного анализа нет способа проверки статистической гипотезы об адекватности полученных классификаций, что делает метод невозможным для содержательного анализа.

### Метод шкалирования

Не всегда возможно измерить характеристику. То есть субъективная информация (качественная, а не количественная). Например, оценка политиков, артистов, певцов.

В качестве исходных данных для шкалирования (применяется когда характеристики объектов не измеряются непосредственно) используются не сами оценки степени сходства объектов, а результаты их ранжирования (неметрические методы).

Например, нет сведений, какая кафедра в МГТУ лучше. Да, у нас есть количественные характеристики по количеству и качеству защитившихся, но нет оценки других.

## Дисперсионный анализ (вариации)

Analysis of variance. Анализ эффектов. Анализ вариации в смысле разброса.

Чем больше размах (дисперсия), тем лучше он для оценки ряда (то есть захватываем больше различной информации).

Вклад экзогенных и эндогенных факторов принципиален, поскольку внешних может не оказаться.

ANOVA — набор методов.

Метод наименьших квадратов и систематизированный набор правил для проверки статистических гипотез называется **дисперсионным анализом**. То есть регрессия с одной или несколькими переменными.

Метод предложен Фишером в предположении, что полное изменение числового вектора $(y_1, y_2, \dots, y_n)$, выраженное через сумму квадратов отклонений от среднего значения $\underline{y} = \frac{1}{n}\sum_{i=1}^{n} \quad y_i$ разбивается на 3 составляющие, связанные с результатом действия различных классификационных факторов $\sum \qquad (y_i - \underline{y})^2 = A + B + AB$, где:

- A — системные взаимодействия (эффект влияния системы саму на себя),
- B — внесистемные (внешние) взаимодействия,
- AB — эффект взаимодействий внутренних и внешних

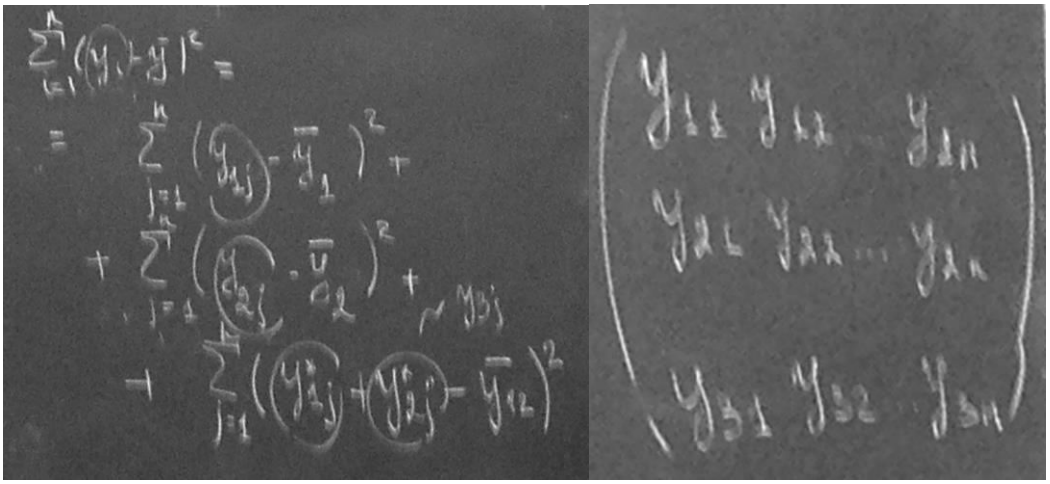
Представим данные в виде матрицы $m \times n$с индексами $i = \underline{1, m};\ j = \underline{1, n}$, тогда общая сумма $\sum_{i=1}^{m} \quad (y_i - \underline{y})^2 = A + B + AB = \sum_{j=1}^{n} \quad (y_{1j} - \underline{y_1})^2 + \sum_{j=1}^{n} \quad (y_{2j} - \underline{y_2})^2 + \sum_{j=1}^{n} \quad (y_{1j} - \underline{y_1} + y_{2j} - \underline{y_2})^2$— общая дисперсия

$y_i$— элементы случайного вектора (скаляр) Относится к у

$\underline{y_1} = \frac{1}{n}\sum_{j=1}^{n} \quad y_{1j}$ — относится к матрице $y_{ij}$

$$y_{3j} = y_{1j} + y_{2j}$$

Получили матрицу неизвестных:



Результаты осреднения влияния факторов A, B, AB.

Производится 3 серии экспериментов, в результате которых находится результат разложения.

Пусть даны три и более обработок, например:

1. $(x_1, x_2, \ldots, x_m)$
2. $(y_1, y_2, \ldots, y_n)$
3. $(z_1, z_2, \ldots, z_{m+n}) = (x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n)$

Введём понятие главного среднего $\underline{z} = \sum_{i=1}^{m+n} \quad \frac{z_i}{m+n} = \frac{m\underline{x}+n\underline{y}}{m+n}$тогда справедливо алгебраическое тождество: $\sum_{i=1}^{m+n} \quad (z_i - z)^2 = \sum_{i=1}^{m} \quad (x_i - \underline{x})^2 + \sum_{i=1}^{n} \quad (y_i - \underline{y})^2 + \frac{mn}{m+n}(\underline{x} - \underline{y})^2$

Первое и второе слагаемые в сумме образует меру изменчивости внутри выборок и называется **внутривыборочной суммой квадратов**. Последний член ряда измеряет различие между выборками и называется **межвыборочной суммой квадратов**.

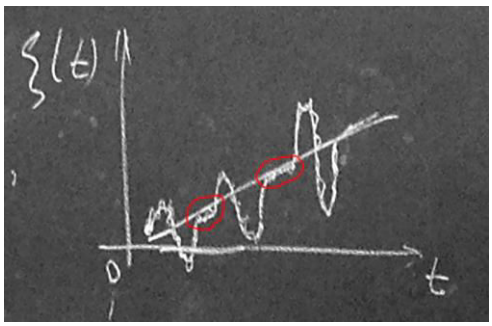
В такой постановке при соответствии двух исследуемых выборок одной генеральной совокупности, проявляющееся различие между ними вызвана случайными флуктуациями. В этом случае внутривыборочное изменение будут заметны.

Если $\sigma^2$— общая ил главная дисперсия выборки z, тогда $(m+n)\sigma^2$ характеризует внутривыборочные изменения.

Для принятия решения заполняется таблица следующего вида.

| | источник изменчивости | сумма квадратов | число степеней свободы |
|---|---|---|---|

| | | | |
|---|---|---|---|
| различие между выборками | | $\frac{mn}{m+n}(\underline{x}-\underline{y})^2$ | 1 (константа) |
| различие внутри выборок | | $\sum_{i=1}^{m}(x_i-\underline{x})^2 + \sum_{i=1}^{n}(y_i-\underline{y})^2$ | m+n |
| полная изменчивость | | $\sum_{i=1}^{m+n}(z_i-z)^2$ | m+n+1 |




# Анализ временных рядов

Лаг — это та часть, где рост совпадает с тенденцией (красным отмечено)

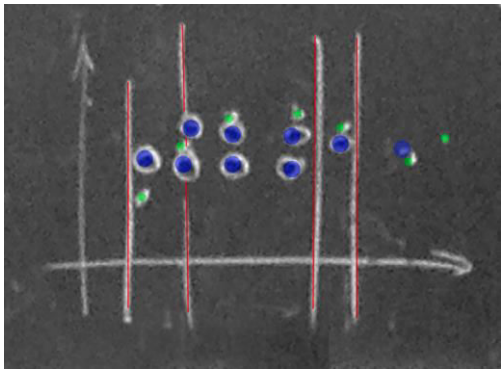
1. Выделение тренда (составляющая 0-го порядка)
2. Выделение периодической составляющей (составляющая 1-го порядка)
3. Анализ выбросов и экзогенные (внешних) воздействий на систему (составляющая 2-го порядка)

*Особенностью является то, что в зависимости от постановки задачи мы можем последовательно выполнять перечисленные этапы, можем реализовывать не все этапы.*

Выделение тренда, при котором используется сглаживание:

## Метод наименьших квадратов

Как правило используется метод наименьших квадратов в предположении, что тренд линеен. Если очевидно наличие нелинейной функции тренда, данные предварительно линеаризуют (например, логарифмируют).

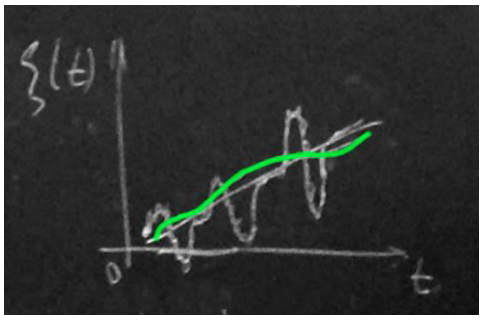


## Метод скользящей среднего

Сейчас узаконились в цифровой обработке данных сжатие методом скользящего среднего:

Если у на подряд n значений. Например, при ширине окна 4 — 4 значения заменяем на 4 средние. Потом переходим к следующим со сдвигом единица (а можно пропустить две). При этом на следующем шаге используются оригинальные значения, а не заменённые. И так далее.

*Если вернутся к предыдущему примеру: кривая, полученная данным методом будет иметь вид, как зеленая линия на графике справа:*




Сглаживание всегда включает некоторый способ сжатия данных. При этом несистематические компоненты временного ряда взаимно гасятся, а систематические дают информацию о тренде временного ряда. Одним из общих методов сглаживания является метод скользящего среднего, в котором каждый член ряда заменяется средним собственного значения и $n-1$соседних членов ряда. В этом случае n называется **шириной окна**.

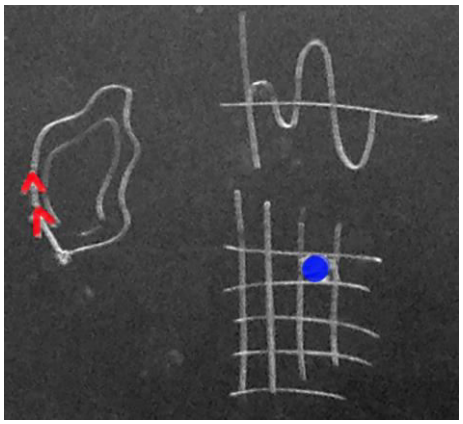
Метод скользящего среднего обладает существенным недостатком: изменяет значения выбросов (срезает их), но не позволяет совсем исключить их. То есть выброс сглаживается, но присутствует.

## Медианный метод

Чтобы избавиться от проблемы метода скользящего среднего, выбираем ширину окна такую, чтобы приходился только 1 выброс в окне. Из попавших в окно значений исключаются минимальное и максимальное и все значения в окне заменяются на среднее оставшихся значений.

---

В результате выделения функции тренда проводится её анализ (достаточно гладкая (непрерывная, не имеет разрывов, дифференцируемая), монотонно возрастающая, разрывы первого и второго рода). Эти знания позволяют делать прогнозы развития процесса на будущие периоды.

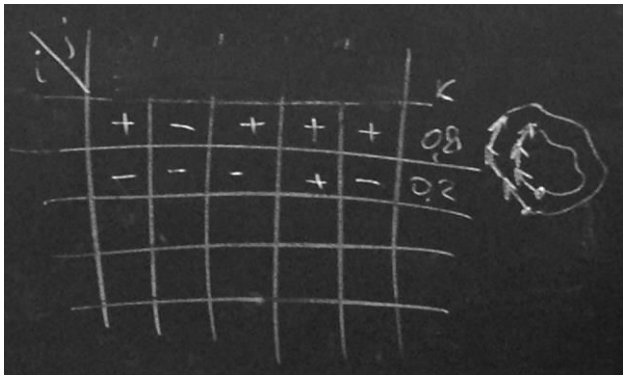


## Периодичность

Периодичность зависимости может быть установлена сравнением значений i-го и j-го наборов значений временного ряда. Распространённым является подход, основанный на построении автокорреляционной функции

(АКФ) (ряд сравнивается сам с собой) и взаимной корреляционной функции (ВКФ). Построение АКФ и ВКФ наследованы из метода контурного анализа, предполагающего замену пары соседних значений на вектор заданной длины. Далее происходит сравнение направлений векторов в предположении, что стартовая точка обхода в одном контуре (периода случайного процесса) фиксирована, в другом — совпадает с началом i-го вектора на i-й итерации сравнения.

После установления периодических изменений исследуемого процесса оценивается продолжительность периода, амплитуда изменения значений и т. д.



Эти подходы предполагают построение кореллалограмм (от корелляции).

### Оценка выбросов и экзогенных процессов на систему

Имея сумму тренда и периодической составляющей временного ряда находят составляющую, определяющую влияние внешней среды на изучаемую систему и выбросы.
Эти сведения позволят дать рекомендации по изменению функционирования системы с целью уменьшения влияния внешней среды на неё.

## Кластерный анализ

Трион в 1939 году предложил данный термин, обобщив несколько подходов в кластреризации объектов. Трион подрумезовал под кластеризацией объединение объектов в группы по схожим признакам (так же понимали и классификацию). Но необходимо отличать понятия кластеризации и классификацией.

Под **классификацией** здесь и далее будем понимать разбиение объектов на группы по заранее известным показателям, полученных на этапе предварительного исследования (обучения). При этом число классов ограничено.

Под **кластеризацией** понимают разбиение множества объектов на кластеры — подмножества, показатели которых заранее неизвестны. Количество кластеров может задаваться произвольным или определяться в процессе кластеризации.

*Задача: абитуриенты могут быть зачислены при А (баллы не менее 200), Б (при наличиии мед. справки) В (при наличии заданного возраста) и у нас всего 2 группы - 1) “Зачислены” 2) “Не зачислены” - мы говорил о классификации.*

Различают **дивизивные** методы кластеризации, предполагающие разбиение всего множества объектов на конечное число кластеров. Также различают **агломеративные** методы, предполагающие объединение нескольких кластеров в один по сходным признакам объектов. При этом кластеров столько (на первом шаге), сколько объектов в выборке.

Кластерный анализ считался не базовым методом статистического анализа, а лишь базовым. Но сейчас кластеризация чаще стала заменять задачи классификации, поскольку более широкие ограничения (можно не определять показатели).

**Постановка задачи кластеризации**
Пусть X - множество объектов, $Y$ - множество номеров кластера. Введём расстояние от объекта до объекта.

Пусть $\rho(x1, x2)$ - расстояние между объектами.

Кластеризация предполагает произвольный выбор на первом шаге различных характеристик кластера. При этом, можно осуществить выбор на основе так называемой обучающей выборке $\{x_1, x_2, \ldots, x_n\} \in X$ эта выборка в свою очередь может быть разбита на несколько стартовых кластеров, в каждом из которых объекты отличаются не более, чем на величину $\rho_0$

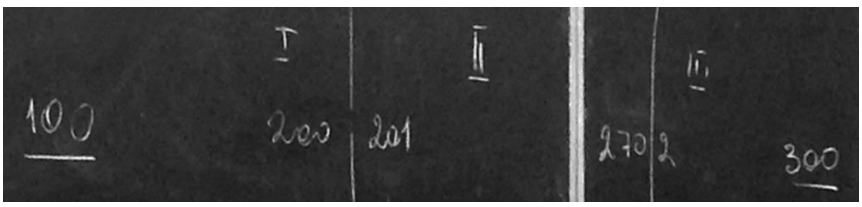
Например: есть абитуриенты с баллом ЕГЭ 100, есть абитуриенты с баллом ЕГЭ 300.

В экспертных системах происходит обучение. То есть решается прямая и обратная задачи.

Объекту $\forall x_i \in X^m$ставится в соответствие номер кластера $y_j$.

*Есть выборка из m элементов (абитуриентов), у каждого своё число баллов. Далее выделяем: первому, второму, третьему - определяем первый кластер ($y_1$), у которых чуть больше баллов — во второй кластер ($y_2$), остальных в третий кластер $y_3$. (предварительная кластеризация).*

*На картинке ниже предварительное разбиение числа абитуриентов по входным баллам с целью оптимизации.*



Следует отметить, что множество Y может быть не до конца известно и может изменяться в процессе кластеризации (а при классификации известно сразу). В том смысле, что могут сливаться, дополняться новыми кластерами…

Решение задачи кластеризации не единственна по ряду причин:

- Не существует однозначно лучшего критерия качества кластеризации.
- Число кластеров неизвестно заранее, а устанавливается по субъективному критерию.
- Различие в стандартизации переменных.
- Результат кластеризации существенно зависит от метрики, выбор которой субъективен и определяется экспертом.

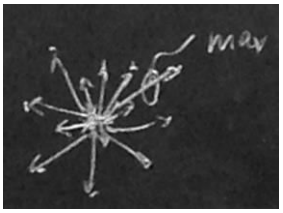
$$\rho_A\,(x_1, x_2) = \sum_{i=1}^{k} \left|x_1^i - x_2^i\right|$$

$$\rho_B\,(x_1, x_2) = \sqrt{(x_1^1 - x_2^1)^2 + (x_1^2 - x_2^2)^2 + \ldots + (x_1^k - x_2^k)^2}$$

$$\rho_C\,(x_1, x_2) = \sqrt[8]{(x_1^1 - x_2^1)^8 + (x_1^2 - x_2^2)^8 + \ldots + (x_1^k - x_2^k)^8}$$

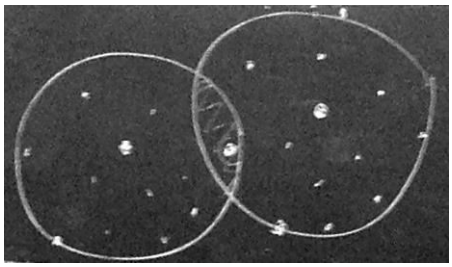
Кластер имеет следующие математические характеристики:

- **центр кластера**. Это среднее геометрическое место точек. Допустим, кластер включает m объектов. Ищем среднее арифметическое по каждому показателю объекта m.
- **радиус кластера**. Это максимальное расстояние объектов до центра кластера. Т.е. все объекты анализируем на предмет расстояния до центра и выбираем максимальную величину. (картинка справа)
- **размер кластера**. Данная характеристика определяет определяется либо по радиусу кластера, либо по среднеквадратичному отклонению объектов от центра кластера.




Формула среднеквадритического отклонения (СКО)::

$$S = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \left(\left(x_i^1 - \underline{x^1}\right)^2 + \left(x_i^1 - \underline{x^1}\right)^2 + \ldots + \left(x_i^1 - \underline{x^1}\right)^2\right)}$$



Если расстояние от объекта меньше радиуса или СКО, то объект относится к данному кластеру. Если условие выполняется для нескольких кластеров, то объект признаётся спорным и, как правило, выделяется в отдельный кластер.

…

## Метод k-средних (пишется видео)

Кластеры представлены в виде центроидов. Центр определяется как центр масст (среднее для каждой координаты).

Алгоритм предполагает построение k кластеров, расположенных на возможно больших расстояниях друг от друга. Выбор числа k производится интуитивно, либо на основе выделенной обучающей выборке.

**Общая идея алгоритма**:

Заданное число k  кластеров сопоставляются так, что средние (центры) кластеров максимально удалены друг от друга. Метод итеративный. Счёт завершается при стабилизации центров кластеров.

**Пример:**
Сейчас рассматривается двумерный пример, но лучше использовать трёхмерный. Предполагается, что будет 3 кластера.

1. Произвольно задали кластеры и посчитали центра.
2. Считаем расстояния от каждого объекта до центра кластера.
3. Если спорный объект однозначно всё же принадлежит определённому кластеру, то назначаем его ему и пересчитываем центр.

Если метод начинает расходится, то просто по-другому разбиваем на кластеры (выбираем центры). Запускаем заново.

Достоинства алгоритма:

1. Простота алгоритма и его реализации
2. Хорошее быстродействие

Недостатки:

1. Слишком чувствителен к выбросам (из-за линейности среднего)
2. На больших объёмов данных бустродействие резко снижается $O(n^2)$. Вынуждены искать расстояние каждого i-ого с каждым j-м
3. Зависимость результата от выбора начальных центроидов. Выбор субъективен, что может привезти к тому, что метод может расходится.

## Метод Ваарда

Метод Ваарда является аггломеративным (на предмет объединения) иерархическим методом. В качестве начальных центроидов выбираются все объекты выборки.

Расстояние между кластерами рассматривается как прирост суммы квадратов расстояний до центров кластеров ($d(X)$), получаемых в результате объединения отдельных объектов $D(X,Y)$.

Характеристика ставится каждому класетру: $D(X,Y) = d(XY) - (d(X) + d(Y))$

$$d(X) = \sum_{i=1}^{N_X} \left| x_i - \frac{1}{N_X} \sum_{j=1}^{N_X} x_j \right|^2$$

В отличие от других методов кластерного анализа в этом методе применяются принципы дисперсионнного анализа (анализируем разброс и, если кластеры далеко друг от друга, то на следующем шаге просто не выполняем сравнение).

На каждом шаге алгоритма объединяются такие 2 кластера, которые приводят к минимальному увеличению целевой функции, то есть внутригрупповой суммы квадратов

Ваард предлагал этот метод для решения конкретных задач. Может потребоваться, когда кластер маленького размера. В целом, количество методов, решаюзих общую такую задачу - много. Если же нужно разделить большую выборку на кластеры малого размера, то лучше использовать метод Ваарда. То есть задача построения иерархической модели или некая классификация.

Метод направлен на объединение близко расположенных кластеров и стремится создавать кластеры малого размера.

### Метод ближайшего соседа (одиночная  связь)

Известен под названием “одиночная связь”. Также является аггломеративным иерархическим. Расстояние в этом случае определяется расстоянием между двумя самыми близкиим объектами

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y), d(x,y) = \rho(x,y)$$

Метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких по единственному признаку элементов. В других случаях предпочтительнее создание кластеров облачного типа (где элементы скомпонованы наоборот). С этой целью применяется **метод удалённого соседа**.

### Метод удалённого соседа

Тоже иерархический аггломеративный метод (метод полной связи).

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y), d(x,y) = \rho(x,y)$$

Это альтернатива цепочным кластерам.

## Общее положение теории планирования эксперимента

Родоначальник — Роберт Фишер 1918г.

На акробиологической станции нужно было провести серию экспериментов. Выдено 3 делянки, с какими-то растениями. Было 3 удобрения, чтобы выяснить, как это влияет на расстения. Урожай 1 раз в год.

Всегда, когда говорим о планировании экперимента - говорим о большом количестве экспериментов. Нужно предложить конечный вариант прогона, который покажет, справедлива ли построенная модель.

Обобщение результатов Фишера было сделано в монографии Design of Experiments, 1935. Далее были получены разными математиками частные результаты, которые Фишер обобщил в 40е годы.

Отличительной особенностью планирования эксперимента является проведение конечного числа опытов по заданной программе. Фишер предложил уменьшить количество экспериментов посредством построения факторных планов (сокращаем количество факторов), а Йетс разработал вычислительную схему факторного эксперимента.

В 1945 год Финни предложил дробные факторные планы, что позволило оптимизировать эксперименты в промышленности.

В начале 40х годов Хотелинк предложил находить экстремум по экспериментальным данным с использованием степенных представлений и градиентной функции. Это позволило производить оптимизацию планов.

В 1947 году Фридман и Севидж предложили итерационный подход к экспериментальному определению экстремума.

Недостаточно проработанной оставалось фаормализация объекта исследования. В связи с чем сначала в кибернетике (науке об управлении) появилось понятие чёрного ящика — замена формального описания модели.

В 1951 году Бокс и Уилсон сформулировали и довели до практических рекомендаций идею последовательного экспериментального определения оптимальных условий проведения процессов.

В 1954-55 годах революционное распространение теории планирования эксперимента на описания физико-химических экспериментов.

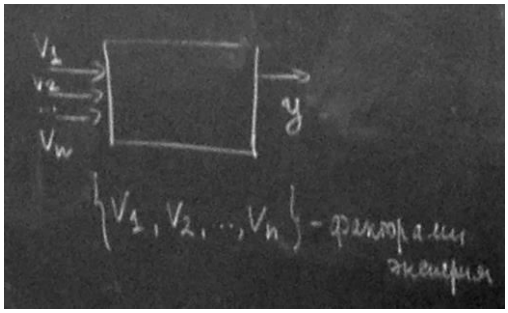
В 60е годы развивается экономическое планирование.

В связи с лавинообразным ростом информации в разных предметных областях и развитию вычислительной техники в 70е годы активно развивается теория планирования вычислительного эксперимента.

Теория планирования эксперимента охватывает практически все встречающиеся варианты исследования объектов. Типовыми являются следующие задачи:

- поиск значений параметров системы, обеспечивающих достижение показателя качества исследуемого объекта при заданных ограничениях
- приближённое аналитическое описание функции связи, показатели качество объекта с параметрами системыпо результатам эксперимента
- оценка дифференциального влияния уровней параметров системы на показатель качества объекта
- испытание образцов техники
- отсеивающие эксперименты (выявление параметров, не влияющих на качество прибора)
- задачи адаптивного планирования (эксперимент проводится над системой, которая постоянно меняет условия работы)

2013-12-21 (Последняя лекция)

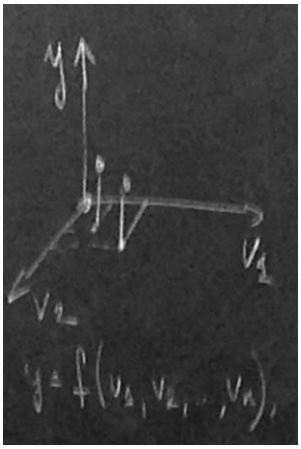



## Метод ближайшего соседа (одиночная связь)

Имеется чёрный ящик. На входе у нас есть набор характеристик, $\{V_1, V_2, V_3, \ldots, V_n\}$— факторы эксперимента.

Планирование эксперимента заключается в построении гиперкуба (размер больше 3). Можем представить в виде обычного куба, если у нас 3 параметра (но это скорее не куб а параллелограмм). Двухфакторный анализ представлен на рисунке справа (то есть то, что мы можем представить на двух осях факторы, а на третьей — значение)

$$y = f(v_1, v_2, v_3, \ldots, v_n)$$



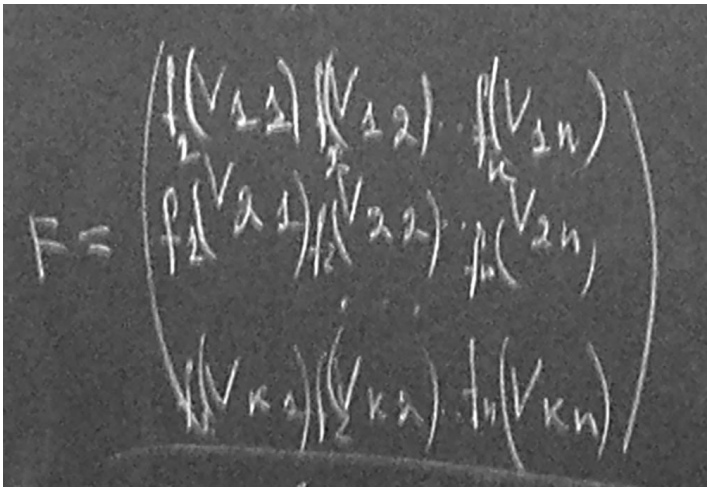

Теория планирования эксперимента изучает **активные** эксперименты, т.е. возможно целенаправленное изменение значение факторов в заданных диапазонах. Назначение диапазонов изменения факторов приводит к построению гиперкуба факторов.

Факторы могут быть **качественными** и **количественными**. Качественный анализ, как правило, квантируют, приписывая им числовые значения

Можем сказать, что множество факторов, таким образом, можно

поставить в соответствие геометрическое понятие факторного пространства. Совокупность конкретных значений факторов определяет точку, соответствующую конкретному испытанию.

$f_1(v_{11})\ f_2(v_{12})\ \dots\ f_n(v_{1n})$



$$F = f_1(v_{21})\ \ f_2(v_{22}) \dots f_n(v_{2n})$$

…

$$f_1(v_{k1})\ f_2(v_{k2})\ \dots\ f_n(v_{kn})$$

Всего — k испытаний. F — матрица плана

$$y = b_0 + (b_1x_1 + b_2x_2 + \dots + b_nx_n) + b_{12}x_1x_2 + b_{13}x_1x_3 + \dots + b_{23}x_2x_3 + \dots + b_1^2{x_2}^2 + \dots + \varepsilon$$

, где $\epsilon$— несущественный остаток (т.е. в остатке всё, что мы не учли)

Пример: анализ запросов к базе данных (допустим, к электронному университету МГТУ). В этом примеры факторов::

- интенсивность запросов
- скорость передачи данных по каналу
- объем доступной памяти на сервере (дисковое пространство)

Кроме того, на объект воздействуют возмущающие факторы, являющие случайными и не поддающиеся управлению. Эти факторы описываются случайной составляющей $\epsilon$.

Сколько будет испытаний - столько будет откликов $(y_1, y_2, \dots, y_n)$, и столько же будет $\epsilon$.

Теория планирования эксперимента (ТПЭ) предполагает построение регрессии вида $y = b_0 + (b_1x_1 + b_2x_2 + \dots + b_nx_n) + b_{12}x_1x_2 + b_{13}x_1x_3 + \dots + b_{23}x_2x_3 + \dots + b_1^2{x_2}^2 + \dots + \varepsilon$ с учётом переопределения системы (k испытаний в эксперименте, k откликов, k значений случайных воздействий *(от испытаний к испытанию может меняться),* k * n - значение фактора).
Для ограничения количества испытаний (k) необходимо построить критерий оптимальности, позволяющий получить оптимальный план эксперимента.

## Критерий оптимальности планов и их типы

В настоящее время используется более 20 критериев оптимальности. Критерий оптимальности делят на две основные группы:

1. К первой группе относятся критерии связанные с ошибками коэффициентов модели $b_0, b_1, \ldots, b_n$.
2. Ко второй группе относят критерии, связанные с оценкой ошибки (ошибок) поверхности (поверхностей) отклика.

Поверхность отклика (сетка на картинке слева):

Критерии первой группы направлены на выделение доминирующих (наиболее значимых) факторов на начальных этапах решения задачи.

*Допустим, было у Фишера 3 типа удобрений, одно из которых содержит, например, краситель. Фишер провёл три исследования, и выявил, что краситель одного из удобрений незначительно влияет на эксперимент.*

Геометрическое толкование свойств ошибок коэффициентов связано с эллипсоидами их рассеяния, определяемого математическим ожиданием функции фактора (в частном случае самого фактора) и дисперсией значения ошибок.

**Критерий d оптимальности** соответствует минимуму произведения всех дисперсий ошибок коэффициентов полинома $b_0, b_1, \ldots, b_n$.

---

Пример задачи:

Например, есть какая-то переопределённая система: Рассматриваемая модель:

$$\underline{E} \in R^2, \underline{E} = (\epsilon_1 + b_0,\ \epsilon_2 + b_0, \ldots, \epsilon_n + b_0)$$

$$\underline{Y} \in R^2$$

Проводим k испытаний

$$y = b_0 + b_{1j}x_{1j} + b_{2j}x_{2j} + \ldots + b_{nj}x_{nj} + \epsilon, j = \underline{1, k}$$

Неизвестные: $b_0, b_{1j}, \ldots, b_{nj}, \epsilon$

$$x_{11}\ x_{12}\ \ldots\ x_{1n}$$

$$X_{kn} = x_{21}\ x_{22}\ \ldots\ x_{2n}$$

…

$x_{k1}\ x_{k2}\ \dots\ x_{kn}$

$$\underline{B} \in R^n, B = (b_1, b_2, \dots, b_n)$$

$$\underline{X_{kn}} \cdot \underline{B_{n,1}} + \underline{E_{k,1}} = \underline{Y_{k,1}}$$

Не знаем, сколько экспериментов нужно провести. Проводя k испытаний, для каждого коэффициента получаем выборку. Нужно, чтобы дисперсия выборок каждых коэффициентов была минимальной. После достижения порога, моем остановиться по количеству экспериментов.

---

Выбор критерия зависит от задачи исследования, так, при поиске оптимальной функции отклика удобнее использовать критерий d - оптимальности. А при изучении влияния отдельных факторов на поведение объекта критерий **e - оптимальности** (план, в котором максимальная дисперсия коэффициентов должна быть минимальна). Критерий **d - оптимальности** вычислительно сложен, поэтому используют его упрощение: критерий **a - оптимальности** (предполагает минимизацию суммарной дисперсии всех коэффициентов).

Критерий второй группы используется при решении задач описания поверхности отклика с целью определения ограничений на значение факторов. Основным здесь является критерий **g - оптимальности**. Данный критерий предполагает минимизацию максимальной ошибки построения отклика.

Среди всех возможных критериев, исследователи выделяют критерии, позволяющие получить планы с заданными характеристиками.

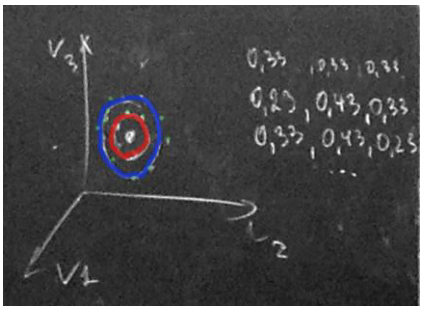
Среди всех классов планов, выделяют ортогональные и ротатабельные планы.

**Ортогональным** называется **план**, для которого выполняется условие парной ортогональности (векторное произведение равно нулю) столбцов матрицы плана. При ортогональном планировании коэффициенты полиномов определяются независимыми друг от друга, вычёркивание или добавление слагаемых в функции отклика не изменяет значение основных значений полинома.

Использование **ротатабельных планов** обеспечивает для любого направления от центра эксперимента равнозначность точности оценки функции отклика на равных расстояниях от центра эксперимента.

Мы хотим получить определённую дисперсию для задач. То есть решаем задачу, исходя из этих заданных требований.

Дисперсия ошибок не должна изменяться в пределах одного кольца (красного, синего итд. - рисунок справа).



Самостоятельно к экзамену:

- определение полного факторного плана
- определение дробного факторного плана

# Системы массового обслуживания

Теория массового обслуживания как задача рассматривалась учёным Эрлангом (швед, статистик). Он работал простым сотрудником телефоннной станции. Было необходимо упорядочить её работу.

- Нужно заранее рассчитать количество потенциальных пользователей в сети в зависимости от числа используемых устройств
- Рассматривались системы с ожиданием и системы без ожидания. Пауза для прерывания другого разговора или сообщения о количестве времени и стоимости после разговора.

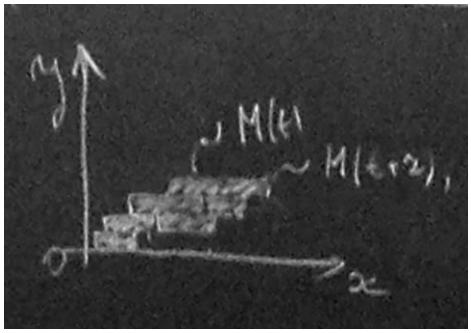
Цель упорядочить привела к структуризации предектной области и выявлению новых возможностей — очереди ожидания.

В случае системы без ожидания звонок терялся. В случае с ожиданием, помещался в очередь ограниченной или неограниченной длины.

**Поток однородный**, если моменты поступления заявок влияют на решение задачи независимо от деталей каждой конкретной заявки. То есть все заявки равноправны.

**Поток без последействия** - если число событий любого интервала времени $(t, t+x)$не зависит от числа событий на другом непересекающемся интервале времени $(t^*, t^*+x^*)$.

Пример таких телефонных звонков: звонки на Новый год



**Потоки с последействием** — когда зависимость есть.

**Стационарный поток** — если вероятность появления n событий на интервале времени $(t, t+x)$ не зависит от времени t, но зависит от длины участка x.

**Простейший поток** —- однородный стационарный поток без последействия.

Простейший поток описывается законом Пуассона.

$\lambda = \frac{M(x)}{x}$—интенсивность заявок в потоке

$\lambda(t) = \lim_{x \to 0} \frac{M(t) - M(t+x)}{x}$, где $M(x)$ — теоретическое матожидание

Формула Литтла (не зависит от типа потока: простейший или нет): $N = \lambda T$