1. Основные задачи математической статистики

Изучение математических моделей случайных явлений или экспериментов в первую очередь занимаются такие науки, как математическая статистика и теория вероятности.

Задачи математической статистики являются обратными к задачам теории вероятности. В теории вероятности после того или иного события/явления требуется рассчитать вероятностные характеристики в рамках данной модели. Моделирование проводится на основе экспериментов, называемых статистическими данными. В ряде случаев, по результатам эксперимента требуется лишь уточнить или модифицировать их.

В задачах математической статистики вероятность того или иного события известна и необходимо получить оценку параметра эксперимента (это могут быть параметры функции связи между двумя показателями объекта, параметры закона распределения случайной величины, в более широком случае - функцию распределения или функцию плотности распределения случайной величины и т.п.).

Как правило рассматривают три задачи математической статистики:

1. **Задача оценки неизвестных параметров по результатам эксперимента.**
   Как правило, нужно найти функцию от результата эксперимента, значения которой являются достаточно хорошей оценкой неизвестного истинного значения параметра.
   $$a - \text{точный параметр}$$
   $$\hat{a} - \text{точечная оценка}$$
2. **Задача интервального оценивания.**
   $$P\{\underline{a} \leq a \leq \underline{a}\} = \gamma$$
   Интервал строится таким образом, чтобы он накрывал неизвестное истинное значение параметра с заранее заданной вероятностью γ - коэффициент доверия.
3. **Задачи проверки стат гипотез.**
   Требуется на основе результатов эксперимента проверить то или иное предположение относительно вида и параметров функции распределения случайной величины и функции плотности распределения.

В математической статистике используется выборочная терминология, основанная на “урновой” схеме. (1) $\{X_1, \ldots, X_N\}$ - **генеральная совокупность**, объёмом N. Этот набор может иметь бесконечную размерность. Из генеральной совокупности выбирается в свою очередь набор (2) $\{x_1 \ldots x_n\}$, $n \leq N$ и набор (2) называется **выборкой из генеральной совокупности** (1), объёма n. Выборка может производиться с возвращением и без него. Если выборка производится с возвращением, то случайные величины в них независимы. В противном случае - зависимы. С “возвращением” тождественно равно независимой

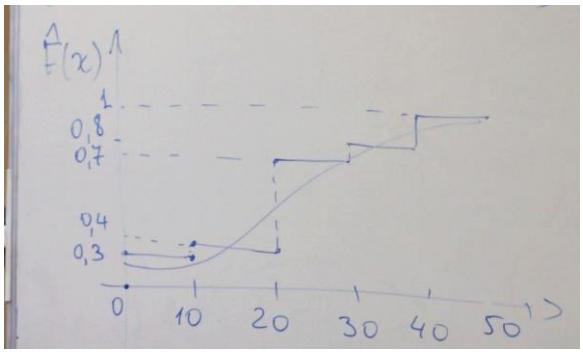
повторной выборке объема n. Терминология сохраняется и в случае с бесконечной генеральной совокупностью.

Числа выборки (2) обычно располагают в порядке убывания или возрастания $(3)\ \{x^{(1)},\dots,x^{(n)}\}$. Набор (3) назвается **вариационным рядом**. Чаще всего, в практических задачах анализируется именно вариационный ряд.

Имперической функцией распределения, построенной на основе выборки (3) (хотя возможно и по выборке (2)), называется функция $\hat{F}(x) = \frac{r(x)}{n}$, n - общее число элементов выборки, $r(x)$- количество элементов $x_i \le x$.

Пример:

$$\{0,0,9,16,21,24,29,37,42,48\}$$



Для моделирования требуется теоритическая функция рапсределения случайной величины x, которая может быть оценена по эмпирической функции распределения.
$sup_{x,n\to\infty}(|F(x) - \widehat{F_n}(x)|) \to 0$. (По теореме Гливенко-Кантелли, теоретическая и эмперическая должная совпадать).

По эмперической функции распредления строят эмперический или выборочный момент.
$\underline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$- выборочное среднее, выборочный аналаог начального момента или мамента ожидания.

$\widehat{\delta^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \underline{x})^2$ - выборочная дисперсия

$\hat{S} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \underline{x})^2}$ - выборочное СКО (среднеквадратичное отклонение).

$\mu_{\xi,a} = (\frac{1}{n}\sum_{i=1}^{n} (x_i - a)^r)^{\frac{1}{r}}$-выборочный момент порядка r.

В ряде случаев требуется оценить разброс результата $R_n = |x^{(n)} - x^{(1)}|$ как размах выборки.

2.Точные оценки параметров

Пусть имеется некоторая случайная величина ξ с функцией распределения $F(x,\theta), f(x,\theta)$ - плотность распределения.

$$\xi \to F(x,\theta_1)$$

…

$$\xi \to F(x,\theta_n)$$

Совокупность распределения даёт параметрическое семейство распределений, в котором θпринимает различные значения.

Вводят функцию от результатов наблюдений (4) $\varphi = \varphi(x_1,\dots,x_n)$, называемую статистикой. Задача построения точечной оценки параметра θ сводится к нахождению значения статистики, такой что $\hat{\theta} = \theta(x_1,\dots,x_n)$: $sup_{n\to\infty}|\hat{\theta} - \theta| \to 0$.

Необходимо установить эффективную оценку, рекомендуемую в качестве результата.

2.1.Свойство оценок

$\hat{\theta} = \theta(x_1,\dots,x_n)$.

Известно, что для экпоненциального распределения $\lambda = 1/\underline{x} = 1/(\frac{1}{n}\sum_{i=1}^{n} \quad x_i) = n/\sum_{i=1}^{n} \quad x_i = \frac{n}{x_1+\dots+x_n}$.

Оценка $\hat{\theta}$ является **несмещённой оценкой** параметра θ, если её математическое ожидание совпадает с теоретической величиной.
Оценка θ является **асимптотически несмещённой**, если $\lim_{n\to\infty} M\hat{\theta}_n = \theta$. Таким образом, на свойство оценок влияет объем выборки.

Если $\lim_{n\to\infty} P\{|\widehat{\theta_n} - \theta| < \varepsilon\} \to 1$, то говорят, что оценка **сходится по вероятности**.

Пусть $\hat{\theta}_n$ асимптотически несмещённая оценка параметра θ, если $\lim_{n\to\infty} S^2(\hat{\theta}_n \ ) \to 0$, то оценка является **состоятельной**.

Таким образом, асимптотическая несмещённость оценки θ, и минимизация разброса значений параметра при $n \to \infty$ обеспечивают состоятельность оценки (теорема приводится без доказательства).

$S^2(\hat{\theta}_n{}^1) = M(\hat{\theta}_n{}^1 - \ \theta)^2 \le M(\hat{\theta}_n{}^2 - \theta)^2 = S^2(\widehat{\theta_n}^2)$, то оценка $\hat{\theta}_n{}^1$ является более эффективной по сравнению с оценкой $\hat{\theta}_n{}^2$.

2.2.Методы построения точных оценок

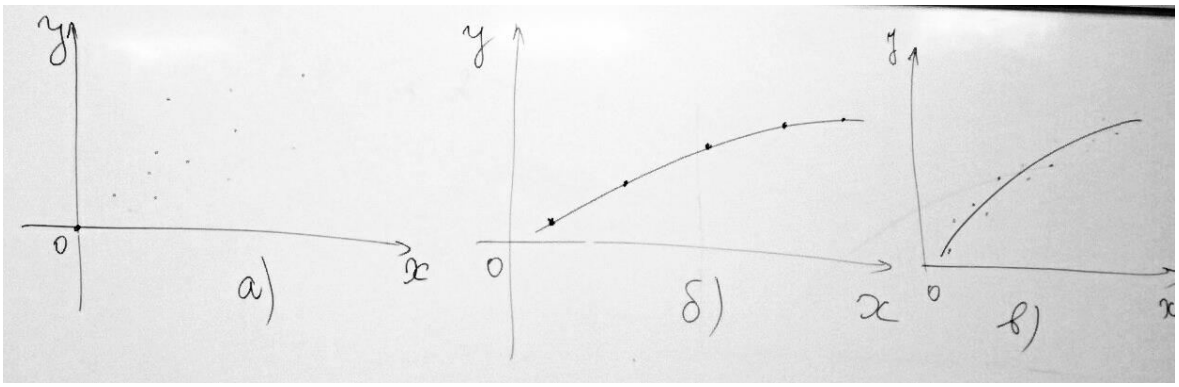
Различаются 2 случая:

- моделирование производят на выборках ξ_1 и ξ_2, наблюдаемых одновременно, в этом случае рассматривают пары точек , характеризующие n объектов в предположении, что характеристика ξ_1, а ξ_2.

$$(x_i, y_j), i = \underline{1,n}$$
$$\xi_1 : \{x_1, \dots, x_n\}$$
$$\xi_2 : \{y_1, \dots, y_n\}$$

Объёмы выборок одинаковы. Как правило для опеределения точечных оценок параметров функции связи между случайными величинами x, y: y=y(x) используют метод наименьших квадратов.

При анализе одновременно наблюдаемых показателей возможны 3 варианта:



В случае **а)** говорят о наличии стохастической связи между двумя случайными переменными. Термин стохастическая связь был введён русским учёным **хз кем**. Стохастическая связь - это такая связь, при которой значению случайной величины $x \in X$ соотвествует(~) одно или более значений $y_1, \dots, y_k \in Y$.

Таким образом, получение случайной величины $y_i, i = \underline{1,n}$ изменяет вероятность появления других значений, но не обеспечивает их появление. Именно поэтому, говорят, что стохастатическая связь не является причиной.

Рассмотрим вариант **б)**. Функциональная связь описывается зависисмостью, в которой случайной величине x из генеральной совокупности X ставится единственное значение y.

$x \in X \sim y \in Y$, т.е. функциональная связь — причина. Предполагается отсутствие ошибок измерения.

Как правило на практике исследователи имею дело с вариантом **в)**:

1. присутствуют ошибки измерения
2. имеет место разброс реализаций относительно некоторой функциональной зависимости в той или иной степени достоверно описывающей входные данные. Такую функцию называют функцией тренда.

В случая достаточно тесной стохастической связяи задача сводится к задаче выделения тренда и его анализа.

Как правило задачу **в)** решают с помощью метода наименьших квадратов.

- Исследуются не наблюдаемые одновременно параметры

$$\xi_1 : \{x_1, \dots, x_n\}$$
$$\xi_2 : \{y_1, \dots, y_m\}$$

2.2.1.Метод моментов

Пусть $x_1 ... x_n$ независимая случайная выборка из генеральной совокупности с функцией распределения $F(x, \Theta)$ и функцией плотности $f(x, \Theta)$, где Θ - параметр распределения.

$F(x, \Theta)$ - теоретические значения ⇒посчитаем (*)$\mu(x) = \int_{-\infty}^{\infty} x\, f(x, \Theta) dx = \frac{1}{n} \sum_{i=1}^{n} x_i \Rightarrow \hat{\theta}$.

(*) позволяет получить оценку неизвестного параметра распределения Θ. В случае многомерного параметра $\underline{\theta}$ распределения используют r первых моментов x_r.

$$\mu_r = \int_{-\infty}^{\infty} x^r f(x, \underline{\theta}) dx$$

$$\underline{x_r} = [\frac{1}{n} \sum_{i=1}^{n} (x_i - a)^r]^b, b = \frac{1}{r}$$

$a = 0$ - момент начальный

$a = \underline{x_r}$ - центральный момент

В общем случае b = 1. Если рост величины момента значителен, то исследователь использует нормализацию $b = \frac{1}{r}$.

Мы берём r первых моментов и решаем систему:

$$\mu_1 = \int_{-\infty}^{\infty} x\, f(x, \underline{\theta}\ ) dx = \underline{x_1} = \hat{\mu}_1$$
$$\mu_2 = \int_{-\infty}^{\infty} x^2 f(x, \underline{\theta}\ ) dx = \underline{x_2} = \hat{\mu}_2$$

...

$$\mu_r = \int_{-\infty}^{\infty} x^r f(x, \underline{\theta}\ ) dx = \underline{x_r} = \hat{\mu}_r$$

В случае двух параметров рассматривают первый начальный момент(математическое ожидание и выборочное среднее) и второй центральный момент(дисперсия и выборочная дисперсия).

В случае оценки функции связи между двумя наблюдаемыми параметрами используют модифицированный метод моментов.

В предположении, что между случайными величинами ξ_1 и ξ_2 имеется достаточно тесная стохастическая случайная зависимость и она может быть описана функциональной зависимостью вида, например: $\xi_1 = \varphi(\ \ \xi_2) = k\ \xi_2$ (линейная зависимость).
В этом случае оценка коэффициентов $\hat{k}$: $\hat{\mu}_{\xi_1} = \varphi(\hat{\mu}_{\xi_2})$

Пример:
$\xi_1 = (100, 200, 300), \hat{\mu}_{\xi_1} = \underline{x}_1 = 200$
$\xi_2 = (10^5, 3 \cdot 10^5, 5 \cdot 10^5)$, $\hat{\mu}_{\xi_2} = \underline{x}_2 = 3 \cdot 10^5$.

$$\hat{\mu}_{\xi_1} = \varphi(\ \ \hat{\mu}_{\xi_2}\ \ ) = k\ \hat{\mu}_{\xi_2} \quad \rightarrow 2 \cdot 10^2 = k \cdot 3 \cdot 10^5 \Rightarrow \hat{k} = \frac{2 \cdot 10^2}{3 \cdot 10^5} \approx 6{,}7 \cdot 10^{-4}$$

или обратно

$$\xi_1 = \varphi(\xi_2)$$
$$\xi_2 = \frac{1}{\hat{k}} \cdot \xi_1 \approx 1{,}5 \cdot 10^3 \xi_1$$

Зависимость между абстрактными случайными величинами:

$$\xi_1 = \alpha {\xi_2}^{\beta}\ (***)$$

Зависимость между конкретными случайными величинами:
$y = \alpha x^{\beta} = \varphi(x)$ - существует стохастическая зависимость в предельном случае между ξ_1, ξ_2

$$\hat{\mu}_{\xi_1} = \frac{1}{n}\sum_{i=1}^{n} \quad y_i = \frac{1}{m}\sum_{j=1}^{m} \quad (\alpha {x_j}^{\beta}\ \ ) = \frac{\alpha}{m}\sum_{j=1}^{m} \quad ({x_j}^{\beta})$$

$$ln\ \hat{\mu}_{\xi_1} = ln(\frac{1}{n}\sum_{i=1}^{n} \quad y_i) = ln(\alpha \cdot \frac{1}{m} \cdot \sum_{j=1}^{m} \quad {x_j}^{\beta}) = ln\alpha - lnm + ln\sum_{j=1}^{m} \quad {x_j}^{\beta}$$

$m \rightarrow n$, $\alpha \rightarrow y_i$

$$ln\ \hat{\mu}_{\xi_1} = ln(\frac{1}{n}\sum_{i=1}^{n} \quad x_i) = ln(y_i \cdot \frac{1}{n} \cdot \sum_{j=1}^{n} \quad {x_j}^{\beta})$$

$$ln\alpha + \beta ln\frac{1}{n}\sum_{j=1}^{n} \quad x_j \Rightarrow \beta A + ln\alpha = A\ (4*)$$

Так как не известны параметры α и β, необходимо построить второе уравнение

$$\widehat{S_{\xi_1}^2} = \frac{1}{n}\sum_{i=1}^{n} \quad (x_i - \widehat{\mu_{\xi_1}})^2 = \frac{1}{m}\sum_{i=1}^{m} \quad (y_i - \widehat{\mu_{\xi_2}})^2 = \widehat{S^2_{\xi_2}}$$

(5*) $\frac{1}{n}\sum_{i=1}^{n} \quad (x_i - \widehat{\mu_{\xi_1}})^2 = \frac{1}{n}\sum_{i=1}^{n} \quad (\alpha x_i^{\beta} - \alpha\widehat{\mu_{\xi_1}^{\beta}})^2 = \widehat{S^2_{\xi_2}}$

Из (5*) получают второе уравнение замыкая систему, получается уравнение такого вида: (6*)

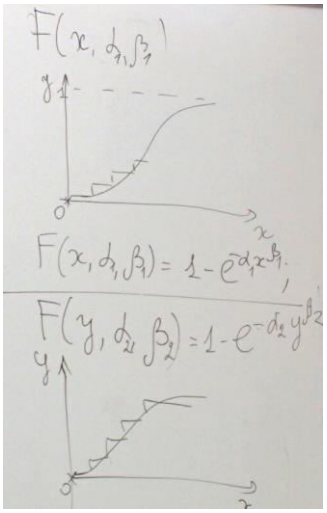
$$1.\ \widehat{\mu_{\xi_1}} = \widehat{\mu_{\xi_2}}$$
$$2.\ \widehat{S_{\xi_1}^2} = \widehat{S_{\xi_2}^2}$$

$\Rightarrow \hat{\alpha}, \hat{\beta}$.

(6*) Это система линейных алгебраических уравнений, во избежание нелинейности (5*) также линеаризуют.

Таким образом, модификация метода моментов оказывается индивидуальна для любого вида функции связи. Кроме того, необходимо обосновать выбор функции связи.

Обоснование вида функциональной зависимости по результатам эксперимента производят на основе анализа выборочных функций распределения. У нас есть что-то с функцией распределения $F(x, \lambda)$.
На рисунке распределение Вейбулла:



$$F(x, \lambda) = 1 - e^{-\lambda x}$$
$$\hat{k} = \frac{\lambda_1}{\lambda_2}$$
$$1 - e^{-\alpha_1 x^{\beta_1}} = 1 - e^{-\alpha_2 \varphi(x)^{\beta_2}}$$

$\alpha_1 x^{\beta_1} = \alpha_2 \varphi(x)^{\beta_2} \Rightarrow \frac{\alpha_1}{\alpha_2} x^{\beta_1} = \varphi(x)^{\beta_2} \Rightarrow \varphi(x) = \frac{\alpha_1}{\alpha_2} x^{\frac{\beta_1}{\beta_2}} \Rightarrow \hat{\alpha} = \frac{\alpha_1}{\alpha_2}, \hat{\beta} = \frac{\beta_1}{\beta_2}$
При этом α_i, β_j- могут быть неизвестными.
Таким образом, подход к точечной оценке параметров функций связи имеет следующие шаги:

1. Анализируются выборки, соответствующие исследуемым величинам, с целью выявления выбросов.
2. Строится выборочная функция распределения вероятностей для обеих выборок. Выборочные функции описываются теоретическим распределением.
3. На основе вида функций распределения анализируется зависимость между исследуемыми переменными (общий её вид). Результатом шага является

обоснование выбора функции связи, параметры при этом остаются неизвестными(но могут иногда быть и изветсными).

4. Анализируется общее количество параметров, рекомендуется приближать многопараметрические функции одно-двупараметрическими.
5. Число уравнений в системе совпадает с числом неизвестных параметров, что позволяет выбрать такое же количество выборочных моментов для обеих выборок и реализовать модифицированный метод момента для получения параметров функции связи между исследуемыми переменными.

Преимуществом стохастического моделирования перед аналитическим является высокая степень формализации к оценке функциональных зависимостей между характеристиками системы, а также алгоритмы обоснования степени влияния того или иного явления на результат исследования. К недостаткам метода можно отнести феноменологический характер модели (построены на результатах конкретного эксперимента) и существенное упрощение видов функции связи на каждом этапе моделирования. Не смотря на субъективность выбора функции связи стохастические модели имеют незначительную вычислительную погрешность, что в некотором смысле компенсирует недостатки (относительную линеризацию), связанные с упрощением.

2.2.2.Метод максимального правдоподобия

Пусть $x_1 \dots x_n$независимая случайная выборка из некоторой генеральной совокупности X с функцией распределения $F(x, \theta)$ и функцией плотности $f(x, \theta)$.

В математической статистике в основе методов лежит выбор функции, зависящей от всех элементов выборки, и называемой статистикой.

В ММП(метод максимального правдоподобия) в качестве функции статистики выбирают $f(x_1, \theta) \cdot \dots \cdot f(x_n, \theta) = L(x_1, \dots, x_n, \theta)$- функция максимального правдоподобия. Оценка максимального правдоподобия находится из соотношения

$$L(x_1 \dots x_n, \hat{\theta}) = max\ L(x_1 \dots x_n, \hat{\theta})$$

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

Если функция L принимает существенные значения, то она логарифмируется: $\frac{dln(L(x_1 \dots x_n, \theta))}{d\theta} = 0 \Rightarrow \hat{\theta}$

В случае многомерного параметра θ, необходимо решить СЛАУ (в общем случае может быть нелинейной системой).

$$\vec{\theta} = (\theta_1, \dots, \theta_p)$$

При фиксированном $x_1 \dots x_n$система будет иметь вид:
(7*)

$$1.\frac{\partial ln\ L(x_1 \dots x_n, \vec{\theta})}{\partial \theta_1} = 0$$

…

$$p.\frac{\partial ln\ L(x_1 \dots x_n, \vec{\theta})}{\partial \theta_p} = 0$$

$\Rightarrow \hat{\vec{\theta}} = (\hat{\theta}_1 \dots \hat{\theta}_p)$, при $\vec{\theta} = (\theta_1 \dots \theta_p)$.

Достоинством такого подхода является получение асимптотически несмещённых и ассимптотически эффективных оценок при $n \to \infty$.

При этом возможен нелинейный вид (7*), что приводит к накоплению вычислительной погрешности.

3. Доверительные интервалы

$$x = \{x_1, \dots, x_n\} \in X_N$$

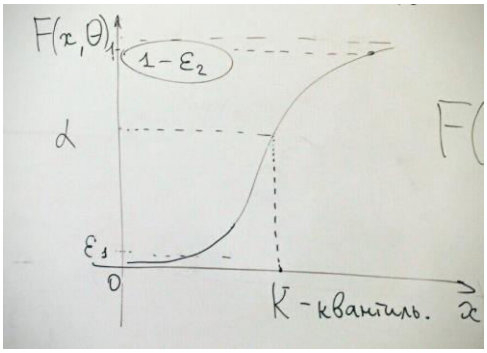
F(x, θ), f(x, θ)

Предположим, что для оцениваемого параметра θ построен доверительный интервал $\theta \in [\theta_-, \theta^-], \theta_- = \theta_-(x_1, \dots, x_n), \theta^- = \theta^-(x_1, \dots, x_n)$. Вероятность попадания в доверительный интервал $P\{\theta_- \le \theta \le \theta^-\} = \gamma$. ($\gamma$= 0.9;0.95;0.99;0.995;0.999)

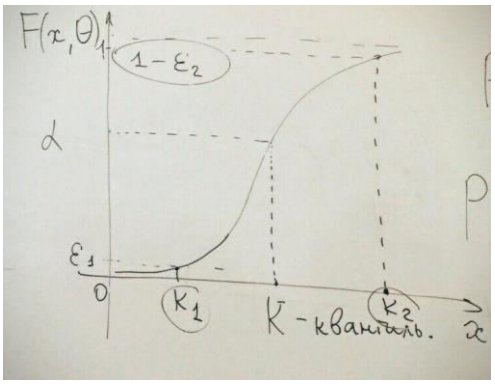
3.1. Методы построения доверительных интервалов

Статистикой метода называется любая функция $T = T(x_1, \dots, x_n)$. Если зависит ещё и от θ, то говорят о центральной статистике, при этом выборка $\{x_1, \dots, x_n\}$ независимая, случайная повторная из генеральной совокупности X_N. Параметр θ - скалярная величина, но в общем случае может рассматриваться как вектор. Функция T является монотонной относительно параметра θ.

Квантиль K уровня α — функция распределения F(x, θ): $F(K) = \alpha$, $F(K) = P\{x \le K\}$.



Зададимся двумя малыми числами ε_1, ε_2 и определим величины k_1 и k_2.



(*)$P\{k_1 < T(x_1, \dots, x_n, \theta) \le k_2\} = \gamma$

$$P\{x \le k_1\} = F(k_1) = \varepsilon_1$$
$$P\{x > k_2\} = 1 - F(k_2) = \varepsilon_2$$
$$F(k_2) - F(k_1) = 1 - \varepsilon_1 - \varepsilon_2$$

Откуда:

$$P\{x > k_1\} = 1 - \varepsilon_1$$
$$P\{x \le k_2\} = 1 - \varepsilon_2$$

В (*) левая граница интервала должна быть замкнутой, то есть

(**) $P\{k_1 \le T(x_1, \dots, x_n, \theta) \le k_2\} = \gamma$

Таким образом $\gamma = 1 - \varepsilon_1 - \varepsilon_2$

$$\gamma = 1 - 2\varepsilon \Rightarrow \varepsilon = \frac{1-\gamma}{2}$$

Данее нижния и верхняя границы θ_-, θ^- доверительного интервала могут быть определены как соответственно минимальное и максимальное значения среди всех параметров θ(среди всех возможных значений), удовлетворяющих (**). Если центральная статистика монотонно возрастает по θ, то границы доверительного интервала находят из системы:

(3*) $\begin{cases} T(x_1, x_2, \dots, x_n, \theta) = k_1 \\ T(x_1, x_2, \dots, x_n, \theta) = k_2 \end{cases}$ (первая тета с нижним, вторая с верхним)

Чаще статистику выбирают таким образом, чтобы она монотонно убывала. В этом случае k_1 и k_2 меняются местами. Тогда получится система:

(4*) $\begin{cases} T(x_1, x_2, \dots, x_n, \theta_-) = k_2 \\ T(x_1, x_2, \dots, x_n, \theta^-) = k_1 \end{cases}$ (аналогично с 3*)

Пример доверительного оценивания для доверительных интервалов (для нормального распредения):

$N(\mu, \sigma)$— исследуемая функция распределения вероятности

Стандартный закон распределения: $N_{0,1} \sim N(0,1)$ — функция распределения веротятн. значений статистики.

Для доверительного оценивания математического ожидания в качестве исходной центральной статистики используется:

(5*) $T = (\frac{x-\mu}{\sigma})\sqrt{n}$.

Задача многопараметрического оценивания на несколько порядков (сложность повышается на порядок) при каждом новом неизвестном параметре вычислительно сложна. В силу чего, исследователи сводят задачу многопараметрического исследования к задаче однопараметрического оценивания. Вид статистики (5*) выбирается исследователем.

Выбранная (5*) обладает следующими свойствами:

- Монотонно убывающая по μ функция, имеющая стандартное нормальное распределение. Вводится ε равная $\frac{1-\gamma}{2}$и квантиль k_ε.

$$\begin{cases} (\frac{\overline{x}-\mu_-}{\sigma})\sqrt{n} = u_{1-\varepsilon} = k_2 \\ (\frac{\overline{x}-\mu^-}{\sigma})\sqrt{n} = u_\varepsilon = k_1 \end{cases}$$

$$\begin{cases} \mu_- = \overline{x} - \sigma\frac{u_{1-\varepsilon}}{\sqrt{n}} \\ \mu^- = \overline{x} - \sigma\frac{u_\varepsilon}{\sqrt{n}} \end{cases}$$

Задача:

За год/помесячно на каждой конечной станции

- Видеокамера (количество входящих) (x_1);
- Количество проданных билетов за тотже период в кассе на конечной станции (за год/помесячно) (x_2);
- Количество прокомпостированных (пробитых) билетов на конечной станции (за год/помесячно) (x_3);

Билеты 3-х категорий (зональные, детские (50%), взрослые).

Проездное (кратковременные, постоянные на месяц).

Цель: Исследование оптимизации прибыли.

Найти: Оценка убытка помесячно {за год} по метрополитену.

Решение:

$$\begin{cases} y = f(x_1, x_2, x_3) \\ u\ ?u(x_1, x_2, x_3), v\ ?v(x_1, x_2, x_3) \end{cases}$$

u, v- ограничения

$$U = \Pi - \mathrm{P}$$

U - убыток

П - планируемый доход

Р - реальный доход

П:

1. задать, например, 100 млн. крон в месяц
2. $\prod = a \cdot x_1^*$, a — средняя стоимость поездки, x_1^* — все воспользовавшиеся за месяц метрополитеном.
3. $\prod = \sum_{i=1}^{k} a_i \cdot {x_1}^{(i)^*}, \sum_{i=1}^{k} {x_1}^{(i)^*} = x_1^*$

$\prod = f_1(x_1^*) = \varphi(x_1)$ ⇒необходимо подтвердить(опровергунть) наличие тесной стохастической связи в предельном случае, описанной зависимостью $\prod = \varphi(x_1)$.

Существует аналитическая зависимость между планируемым доходом и количеством входящих в метрополитен в течение месяца. В этом случае можно предположить, что между доходом и количеством пассажиров, вошедших на конечных станциях, существует тесная стохастическая связь между планируемым доходом и количеством пассажиров, вошедших на конечных станциях.

1. $x_1^* = k \cdot x_1$ — грубая оценка
2. $x_1^* = \xi(x_1) \rightarrow \varphi(x_1)$

…

Возможна грубая оценка (на основе центральной предельной теоремы) общего числа входящих в метрополитен в течение месяца (года) (1). Оценку пассажиров, вошедших на 6 конечных станций умножаем на количество станций ($k = 61$). Альтернативой является оценка стохастической связи (2), что требует проведения натурного эксперимента, достаточно длительного и дорогостоящего. Как правило, проводят ускоренные испытания, снижающие ресурсозатратность эксперимента (сокращение по времени). +Использование иных принципов ускоренных экспериментов (...).

P:

1. Получение прямой информации о доходах метрополитена
2. $P = g(x_2^*) = g(\xi_2(x_2)) = g(x_2^*, x_3^*) = g(\xi_2(x_2), \xi_3(x_3)) = \varphi_2(x_2, x_3)$ — стохастическая зависимость.

В предположении, что между доходом и количеством проданных билетов имеется функциональная зависимость, а между билетами, проданными в метрополитене, и билетами, проданными автоматами, на конечных станциях метро существует стохастическая связь вида: $x_2^* = \xi_2(x_2)$.

Так как билеты могут быть проданы вне метрополитена, а также использоваться в любой день после продажи, связь между количеством проданных билетов во всём метрополитене и количеством прокомпостированных билетов в конкретный день исследования является стохастической. Необходимо подтвердить (или опровергнуть) наличие тесной связи между переменными x_2^*и x_3^*, и оценить зависимость $P = \varphi_2(x_2, x_3)$.

$U(x_1, x_2, x_3) = \varphi_1(x_1) = \varphi_2(x_2, x_3)$, в предположении: $x_1^* = \xi(x_1)$, $x_2^* = \xi_2(x_2)$, $x_3^* = \xi_3(x_3)$.

$$\begin{cases} x_1^* = \xi_1(x_1) \\ x_2^* = \xi_2(x_2) \\ x_3^* = \xi_3(x_3) \end{cases}$$

СМО потребовало бы установления законов распределения, что не доступно при коротких сроках испытаний.

Реализовать процедуру установления фукнции связи между ненаблюдаемыми одновременно параметрами. Протестировать на любых наборах данных (до 20 значений).

4. Основные понятия теории проверки статистических гипотез

Статистической гипотезой H называют любое утверждение относительно функции распределения $F(x)$ случайной величины x, касающееся видов функции распределения и значения её параметра.

Гипотезы H проверяются путем сопоставления выдвинутых предположений с результатами эксперимента, которые в статистиске представляют собой n независимых повторных случайных наблюдений над случайной величиной x.

Различают две постановки задач:

1. Теоретическая функция распредления считается известной, в этом случае проверяемая гипотеза называется простой;
2. Теоретическая функция распредления неизвестна, проверку статистической гипотезы осуществляют для семейства функций распределения, отличающихся значениями параметров (сложная гипотеза).

Статистические гипотезы проверяют с помощью статистических критериев.

H_0 — основная гипотеза (H_0 — нуль-гипотеза)

H_A — альтернативная гипотеза (H_1 — первая гипотеза)

Гипотеза, справедливость которой проверяется в результате эксперимента, называется основной, в зависимости от того, какие альтернативы основной гипотезе возможны в предметной области, формулируют альтернативную (конкурирующую) гипотезу.

Статистический критерий — совокупность правил, позволяющих по полученной выборке принять основную гипотезу и отвергнуть альтернативную или наоборот. Имеется общий принцип построения статистических критериев: $\underline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

1. задаётся некоторая функция $S = S(x_1, \ldots, x_n)$ называемая статистикой критерия
2. Ω; $T_o, T_{кр}, T_и \in \Omega$ (множество всех значений статистики Ω разбивается на $T_o, T_{кр}, T_и$).

T_o — множество принятия решений (соответствует основной гипотезе)

$T_{кр}$ — критическое множество (соответствует альтернативной гипотезе). Если критерий таков, что критическими являются как малые, так и большие значения статистики, то критерий называется двухсторонним, в противном случае — односторонним.

Множество, отделяющее основную гипотезу и альтернативную называют зоной индиффирентности ($T_и$).

3. Если $S \in T_o$, то H_0;
   Если $S \in T_{кр}$, то H_A.

В силу того, что $S = S(x_1, \ldots, x_n)$ является случайной (случайной величиной), событие $S \in T_{кр}$ может произойти как при справедливости H_0, так и при справедливости H_A.

Ошибкой I рода называется возможность принятия альтернативной гипотезы H_A, когда верна H_0.

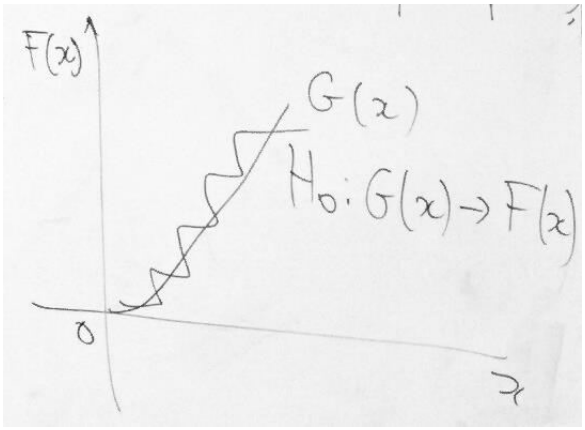
Ошибкой II рода назывется вероятность принятия H_0, когда верна H_A.

$\alpha = P(S(x_1, \ldots, x_n \in T_{кр}) \mid H_0)$ — вероятность ошибки первого рода;

$\beta = P(S(x_1, x_2, \ldots, x_n \in T_o) \mid H_A)$ — вероятность ошибки второго рода.

α и β незначительные величины (в идеале вообще равны нулю) цель исследователя минимизировать эти значения, но одновременно α и β не минизируются (они связаны функционально). Вводят (псевдо-) величину $\gamma = 1 - \beta$, называемую мощностью критерия.

αзадаеётся исследователем (как правило, незначительная величина ~0.01), при этом величина γ максимизируется.
α называют размером критерия. Чем ниже диапазон построения, тем ниже уровень доверия.
β называют уровнем значимости критерия.

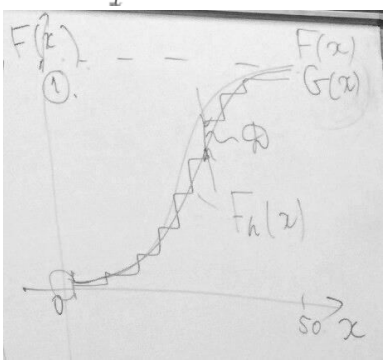


Если известен закон распредления случайной величины, статистические выводы оказываются достаточно точными, в то же время необходимо проверить соответсвуют ли экспериментальные данные этому распределению. Для проверки используют критерий согласия, в частности, критерий Колмогорова-Смиронова и критерий ω^2.

5. Критерий Колмогрова-Смирнова

Критерий основан на статистике Колмогорова. Теоретическая и выборочная функции распредления сравниваются в некоторой равномерной метрике:

(*) $D = \sup_x |G(x) - F_n(x)|$ — статистика Колмогорова.
$D = \max_x |G(x) - F_n(x)|$



$$H\colon G(\cdot) \approx F(\cdot)$$
$$G(x) \; ? \; F_n(x)$$

Критери КС $D \le D_\beta$

Если гипотеза Hверна и при $n \to \infty$: $F_n(x) \to G(x)$, $G(x) \neq F(x)$, то $F_n(x) \to F(x)$ при $D \leq D_\beta$, где D_β выбирается из процентных таблиц точек, как величина на пересечении строки n и β(уровня значимости). При $n > 35$ D_β вычисляется как $D_\beta = \sqrt{-0.5\ ln\beta}, n > 35$.

Статистический критерий для проверки гипотезы H_0 называют состоятельным против альтернативы H_A, если вероятность отвергнуть основную гипотезу (H_0), когда на самом деле верна H_A, стремится к 1 при неограниченном росте n (наблюдений).

Состоятельность критерия Колмогорова-Смиронова означает, что любое отличие распределения выборки от теоретического будет обнаружено, если измерения продолжаются бесконечно долго, что обеспечивается невсегда.

Альтернативой является применение критерия $\omega_n^2 = \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 dF(x)$.

Обычно пользуются эмпирической оценкой $n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} [F(x_i) - \frac{2i-1}{2n}]^2$.

Применение критерия аналогично, величина ω_n^2 (мб даже с н) не должна превышать табулированного порогового значения.

6. Методы многомерного статистического анализа

I. $\xi_1 = \varphi(\xi_2, \xi_3, \ldots, \xi_n)$
II. $\{\xi_1, \xi_2, \xi_3, \ldots, \xi_n\}$

Каждый объект выборки может содержать наблюдения более чем над одной случайной переменной (задача множественной регрессии). Различают две постановки задачи:

1. ξ_1 — зависимая переменная, $\xi_2, \ldots, \xi_n$ — независимые переменные (I);
2. в другом случае — рассмотрим вектор независимых случайных величин, имеющий многомерное распределение (II).

Многомерный анализ - анализ множественных измерений свойств случайной выборки, для анализа которых используется группа статистических методов.
В рамках многомерного статистического анализа ставятся следующие задачи:

- задача стохастического моделирования;
- исследование структуры сложной системы или её модели, в том числе описание поведения системы во времени (изменение тренда, описывающего поведение системы, поиск периодических колебаний(квазипериодических), оценка задержек(лагов));
- задача прогнозирования будущего развития процесса;
- анализ взаимодействий между процессами;
- прогнозирование взаимодействия двух и более процессов;

Для решения этих и аналогичных задач используется анализ временных рядов:

1. методы корреляционного анализа;
2. методы спектрального анализа (ОВР);
3. методы сглаживания фильтрации;
4. методы авторегрессии;
5. методы скользящего среднего;

$y(x) = f(x) + \varepsilon(x)$, $f(x)$-детерминированное значение, $\varepsilon(x)$- случайная составляющая.

Особое место при анализе сложных систем занимают задачи классификации и кластеризации объектов. Пусть имеется совокупность объектов, разбитая на несколько групп (заранее можно сказать какой объект к какой группе относится). Требуется найти группу, к которой относится вновь поступивший объект.

Для решения задач классификации с двумя группами как правило используется дискриминантный анализ. Дискриминантный анализ предполагает построение дискриминириующей функции, аргументами которой являются измеряемые величины,

далее выделяются области, при попадании в которые объект относится к первой или второй группе. Если требуется на разделение на 3 и более групп и/или нет сведений о характиристиках объектов, определяющих группу, используются методы кластерного анализа. Все методы кластерного анализа преполагают идентификацию ядер кластера (измеряемой харакетристики, являются координатами точек в гиперпространстве), оценивается близость вновь прибывшей точки (или любой точки набора) до ядра каждого кластера.
Возможны ситуации, когда характеристики объекта неизмеряемы (возможность качественной оценки лучше (хуже) или количественные оценки неинформативны), тогда используют задачи шкалирования (ранжирования). Такая группа методов называется методами шкалирования.

7. Кластерный анализ

Пусть X множество объектов, Y - множество меток(имён) кластеров. Известна функция расстояния между объектами $\rho(x', x''), x', x'' \in X$. Имеется конечная (обучающая) выборка объектов $X_m \subset X$, $X_m = \{x_1, \ldots, x_m\}$. Требуется разбить множество объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких друг к другу по метрике ρ, при этом сами кластеры удалены друг от друга по метрике ρ. $x_i \in X: x_i \sim y_j$, y_j - метка кластера, $y_j \in Y$. i- бесконечный счётчик. При обучении: $i = 1, \ldots, m$, при кластеризации: $i = 1,2, \ldots$, $j = 1, \ldots, n$ n- задано или вычислено.

Алгоритм кластеризации - это функция $a: X \to Y$, при этом каждому из элементов множества X ставится в соответствие элемент из множества Y.

Решение задачи на основе алгоритма кластеризации принципиально неоднозначно.
Не существует однозначно наилучшего критерия качества кластеризации. Как правило, пользуются эвристическими критериями. Кроме того, возможна кластеризация, по построению сильно зависящая от входных данных. Оптимальное число кластеров при решении задачи как правило не известно и устанавливается на основе некоторого субъективного критерия.

Стандартизация или нормализация переменных приводит к моделированию данных в едином диапазоне значений путём выражения через отношение этих значений к величине, отражающей определённые свойства конкретного признака. Стандартизация производится с помощью экспертной оценки начисления весов или с помощью статистических вычислений отклонений соответствующих переменных. Результат кластеризации существенно зависит от метрики, выбор которой также субъективен и определяется экспертом. Кластер имеет следующие математические характеристики: центр, радиус, среднеквадратичное отклонение (СКО), может быть задан размер

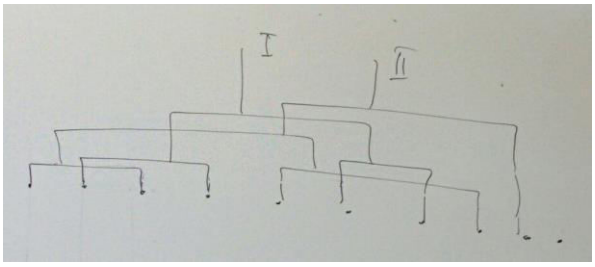
кластера. Центр кластера - среднее геометрическое место точек в пространстве переменных. Размер кластера (радиус) - максимальное расстояние от точек кластера до его центра. СКО - величина квадратично зависящая от расстояния между всеми точками кластера и его центром (на самом деле квадратично зависима дисперсия).

Кластеры не должны быть пересекающимеся, но возникает такое при решении практичкских задач. Объекты, отнесенные двум или более кластерам, называются спорными, решение считается достигнуым, если при выбранной метрике и количестве кластеров спорных объектов нет.

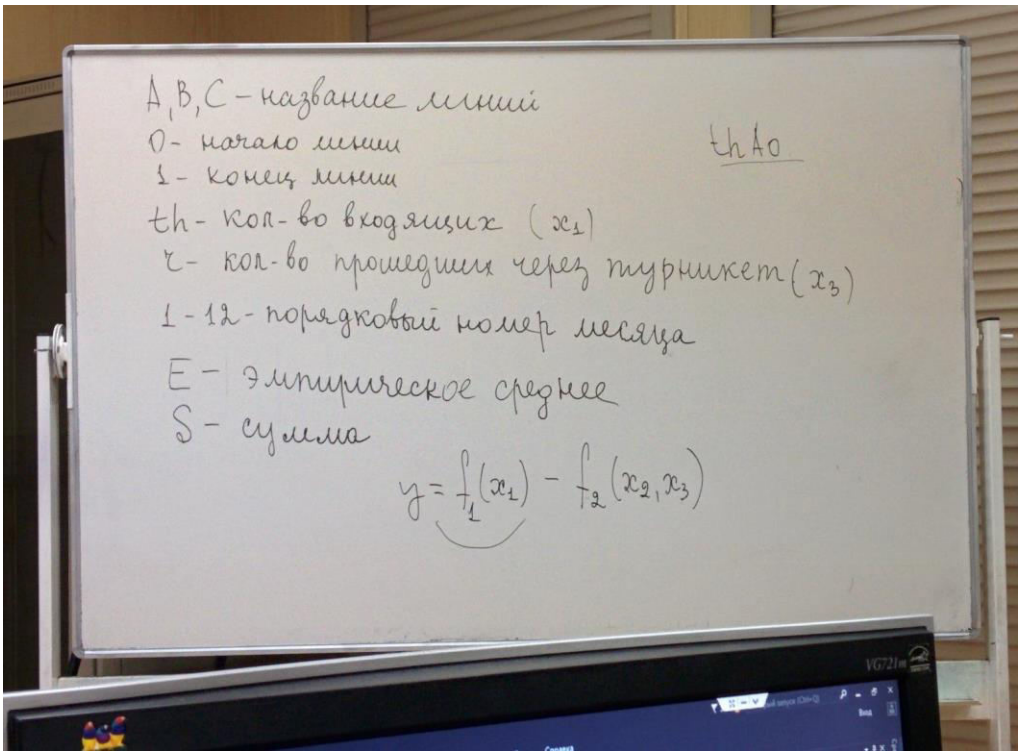
Размер кластера может быть определён либо по радиусу кластера, либо по СКО, либо просто задан. Если часть объектов не может быть кластеризована в связи с ограничением размера кластера, то они также относятся к спорным. Необходимо произвести повторную кластеризацию, вводя дополнительную характеристику - координату. Методы кластерного анализа можно разделить на иерархические и неиерархические.

7.1. Иерархические методы.

Суть иерархии состоит в последовательном объединении кластеров меньшего размера в бОльшие. Или разделение бОльших по размеру кластеров на меньшие. В начале работы все объекты являются отдельными кластерами. На первом этапе наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока объекты не будут объединены в заданное число кластеров. Подход называется агломеративным. На рисунке представлена иерархия дендрограммы:



Сам подход называется агломеративным. Количество кластеров задаётся.
Альтернативой агломеративным подходам является дивизимные методы, то есть в начале работы все объекты принадлежат одному кластеру, на каждом шаге группы расщепляются. Процесс продолжается до достижения заданного числа кластеров.



7.2. Метрики $\rho(x', x'')$

Понятие «расстояния» между объектами отражает меру сходства объектов между собой по всей совокупности использованных признаков.

Пусть расстояние между объектами измеряется скалярной величиной d_{ij}, которая удовлетворяет следующим условиям:

1. $d_{ij} \geq 0$ (неотрицательность расстояния);
2. $d_{ij} = d_{ji}$;
3. $d_{ik} + d_{kj} \geq d_{ij}$(неравенство треугольника);
4. если $d_{ij} = 0$, объекты i и j не тождественны.

В предположении, что $d_{ij} = 0$ и объекты тождественны, говорят о неразличимых объектах.

Вместо термина расстояние в математике применяют термин метрика.

В математике наиболее распространенной и универсальной является метрика Минковского:$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n} (x_i^k - x_j^k)^\lambda}$, n-мерное пространство, n признаков чот там

Разница между количественными характеристиками k-го признака определяется как $(x_i^k - x_j^k)$.

Для $\lambda = 1$: $d_{ij} = \sum_{i=1}^{n} \left|x_i^k - x_j^k\right|$. Модуль в метрике вводится для независимости от знака. Такая метрика носит название расстояние городских кварталов (расстояние Манхэттена).

Для $\lambda = 2$: $d_{ij} = \sqrt{\sum_{i=1}^{n} (x_i^k - x_j^k)^2}$ - расстояние Евклида.

Для $\lambda = \infty$: $d_{ij} = max \ \ |{x_i}^k - {x_j}^k|$ - расстояние Чебышева.

Существуют иные метрики принципиально отличающиеся от метрики Минковского.

$d_{ij} = \frac{\sum_{k=1}^{n} \ |x_i^k - x_j^k|}{\sum_{k=1}^{n} \ |x_i^k| + \sum_{k=1}^{n} \ |x_j^k|}$ - метрика Канберры, или $d_{ij} = \sum_{k=1}^{n} \ \ \frac{|x_i^k - x_j^k|}{|x_i^k| + |x_j^k|}$.

Метрика Канберры используется обычно для данных, измеряемых в диапазоне $\underline{0,1}$.

На основе опыта исследователя можеть быть выбрана любая известная или вновь предложенная метрика. Выбор зависит от количества признаков, диапазона изменяемых данных и функции их изменения.

7.3. Оценка качества кластеризации

Оценка качетсва кластеризации может быть следующим образом:

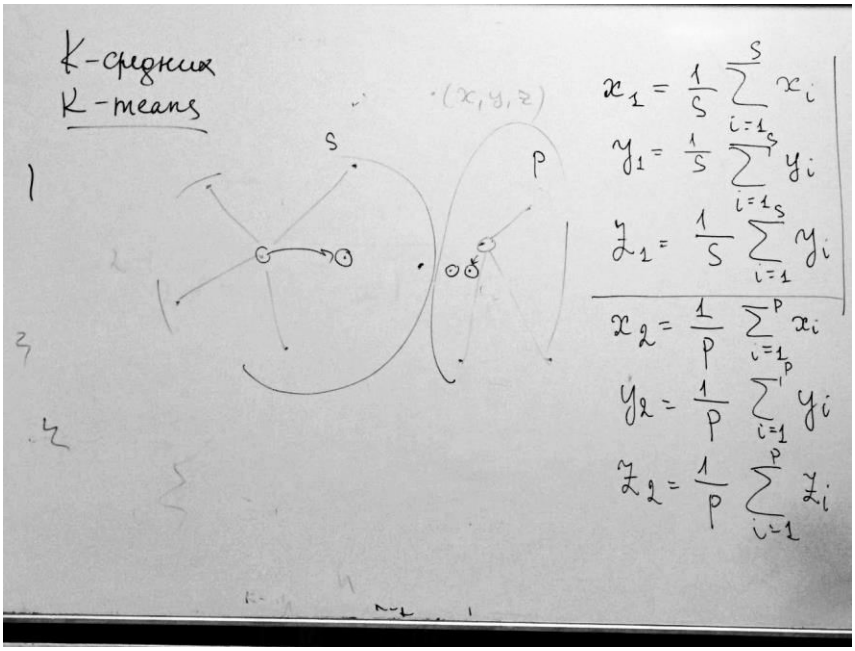
1. На тестовых данных;
2. Установление контрольных точек и проверка на полученных кластерах;
   Контрольные точки интуитивно выбираются исследователем (признаки отчисляемого и не отчисляемого после первой сессии студента). При этом, набор контрольных точек ограничен, а выбор субъективен.
3. Добавление в модель новых переменных.

Если при введении дополнительного признака кластеризация стабильна, то это свидетельствует высоком качестве кластеризации.

Создание и сравнение кластеров с использованием различных методов (различные методы продуцируют разные наборы кластеров, но если в целом результаты сходны, то можно говорить о высоком качестве кластеризации).

7.3.1. Метод К-средних

Метод K-means (K-средних) - как правило его описывают как иерархический. Кластер представлен в виде центройдов, являющихся «центром масс» всех объектов, входящих в кластер. В отличии от классических иерархических подходов, которые не имеют предварительного предположения относительно количества кластеров, в алгоритме К-средних необходимо проверить гипотезу о наиболее вероятном количестве кластеров. Алгоритм К-средних строит K кластеров, расположенных на возможно больших расстояниях друг от друга, то есть кластеры должны бытькак можно более удалены друг от друга. Общая идея: фиксированное число кластеров сопоставляется кластерам таким образом, что средние в кластере для всех переменных близки, но значительно удалены от средних другого кластера.



Из-за смещения центра масс кластера, на каждом новом шаге точки могут перемещатся из кластера в кластер. На i-ом шаге процесс стабилизируется, что свидетелствует о завершении кластеризации. Возможна ситуация зачикливания алгоритма, в этом случае необходимо определить условие окончания счета, если переход повторился, то кластеризацию можно считать завершенной.

Стоимость кластера — функция, зависящая от числа объектов. Если в каждом k-ом кластере S элементов, то стоимость — $c(S_k)$.

Ставится задача минимазации суммы стоимости кластеров по всем кластерам:

$min \sum_{i=1}^{k} \quad c(S_i)$, k— количество кластеров.

Функция стоимости субъективна, выбирается исследователем из соображений возможности её минимизации.

Как правило, минимизируют суммы квадратов расстояний до центров кластеров (для нормализованных данных), либо суммы дисперсий.

Достоинства алгоритма К-средних:

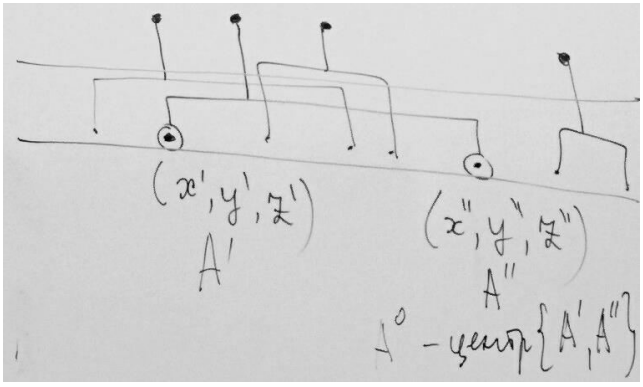
- простота использования;
- быстродействие;
- простой алгоритм выбора начального приближения центров кластеров (например, первые k объектов).

Недостатки алгоритма К-средних:

- алгоритм очень чувствителен к выбросам;
- алгоритм плохо работает на больших объемах данных;
- выбор в качестве начальных центроидов первых kцентроидов негативно влияет на точность результатов из-за ошибочной классификации спорных оъектов (и время).

7.3.1. Метод Варда

Метод Варда является иерархическим агломеративным, в качестве начальных центроидов выбираются все объекты выборки. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний до центра кластеров, на каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции.



A^0 — центр $\{A', A''\}$,

$\rho(A', A^0) = \sqrt{(x' - x^{(0)})^2 + (y' - y^{(0)})^2 + (z' - z^{(0)})^2}$,

$\rho(A'', A^0) = \sqrt{(x'' - x^{(0)})^2 + (y'' - y^{(0)})^2 + (z'' - z^{(0)})^2}$.

Фактически для оценки расстояний между кластерами используются методы дисперсионного анализа, минимальное увеличение функции стоимости соответствует требованию к минимизации внутригрупповой суммы квадратов, то есть метод Варда обеспечивает объедение близких (сходных) объектов при удаленности самих групп.

Условием окончания счета является достижение расстояния между кластерами больше заданного, если оно недостижимо — счет останавливаются по достижению заданого количества кластеров.

Расстояние вычисляется по следующей формуле:

$D(X, Y) = dev(XY) - (dev(X) + dev(Y))$,

XY — слитые кластера,

$dev(X) = \sum_{i=1}^{n} \quad (|x_i - \frac{1}{N_x}\sum_{j=1}^{N_x} \quad x_j|^2)$ — дивергенция.

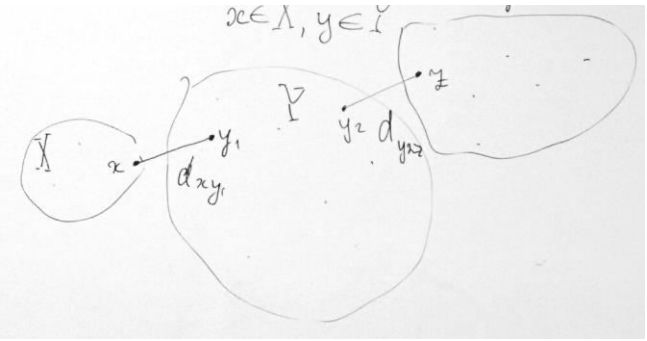
N_x-количество в кластере X .

n - количество рассматриваемых точек.

7.3.2. Метод ближайших соседей

Метод классифицируется как иерархический агломеративный, расстояние между кластерами оценивается как:

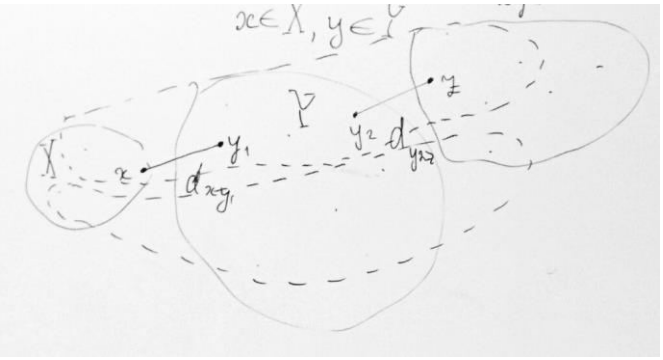
$D(X, Y) = min_{x \in X, y \in Y}(x, y) = d_{xy}$.



Как правило, используется в задачах, где объекты связаны иерархической связью (например, родовидовые связи в биологии).

Достоинство метода: позволяет выделять кластера сколь угодно сложной формы, при условии, что различные части кластера соединены цепочкой. Такие класетры называют цепочечными или волокнистыми.

Недостаток: структура кластера по сути определяется случайными объектами.



7.3.3. Метод наиболее удаленного соседа

В прямом переводе метод иногда называют методом полной связи.

Относится к иерархическим агломеративным, расстояние между кластерами определяется формулой:

$D(X,Y) = max_{x \in X, y \in Y}(x,y) = d_{xy}$.

Расстояние между кластерами определяется наибольшим расстоянием между любыми двумя объектами в различных кластерах.

Метод хорошо работает и создает кластеры, приближенные к гиперсферам, если объекты происходят из разных «рощ». Если кластеры имеют естественный цепочечный вид, то метод использовать не стоит.

Таким образом, можно рекомендовать использовать метод ближайщего соседа при описании иерархический структур, метод удаленного соседа — для описания данных, гарантированно определяющих удаленные кластеры при тесной связи объектов внутри кластера. Метод Варда используют с той же целью, что и метод удаленного соседа, но

работает с облаком точек. В случае, когда предварительной информации о данных нет, используют неиерархический подход, в частности, метод К-средних.

/////
Приближенное вычисление

$$a \approx 1{,}2$$
$$b \approx 1{,}1$$

$$\begin{cases} x > d \\ d > ax + b \end{cases}$$

a, b, d- заданы.

7.3.4. Многомерный дисперсионный анализ

Пусть для каждого из n объектов измеряются k переменных. Тогда результаты измерений могут быть представлены таблично (n - столбцов, k - строк). Вектор столбец $\underline{x_i}$ соответствует i-ому объекту. Для каждого такого объекта можно построить линейную регрессионную модель. При объединении столбцов в матрицу можно составить многомерную линейную обобщённую модель.

$$\underline{X_{1,n}} = A_{1,k-1}X'_{k-1,n} + \underline{\epsilon_{1,n}}$$

A - коэффициенты?
$\underline{\epsilon}$ - вектор погрешностей включает свободные члены и погрешности измерений.
Таким образом, задача сводится к оценке вектора неизвестных коэффициентов и вектора ошибки.
В общем случае решается переопределенная система линейных алгебраических уравнений, то есть количество уравнений должно превышать количество объектов.
Пример?
Пусть исследуется 16 факторов, влияющих на успеваемость студентов. Требуется построить модель многомерного дисперсионного анализа, позволяющую классифицировать студентов на успевающих/задолжников/неуспевающих. Для задачи был анкетирован 461 студент, из которых 33 являлись успевающими, 29 - неуспевающими, 399 - задолжниками.

$$A^T_{k-1,1}\,\underline{X}_{1,n} = A^T_{k-1,1} \cdot A_{k-1,1} \cdot X'_{k-1} + A_{k-1,1} \cdot \underline{\varepsilon}_{1,n}$$

$$\underline{Y_j} = \{Y_{j,1}, Y_{j,2} \dots Y_{j,461}\}$$
$$\underline{Y_i} = \{Y_{1,i}, Y_{2,i} \dots Y_{16,i}\}$$

Следовательно, матрица описывающая эксперимент имеет размерность 16*461. Каждому j-ому фактору поставим в соответствие $\beta_j = \{\mu_j, \beta_{j,1}, \dots, \beta_{j,461}\}$. Тогда j-ая строка матрицы будет иметь вид: $Y_j = \{1, Y_{j,1}, \dots, Y_{j,461}\}$.

$$\begin{cases} \beta_j = \{\mu_j, \beta_{j,1} \dots \beta_{j,461}\} \\ Y_j = \{1, Y_{j,1}, Y_{j,2} \dots, Y_{j,461}\} \end{cases}$$

$$\begin{cases} \beta_j = \{\mu_j, \alpha_{j,1}, \alpha_{j,2}\} \\ Y_j = \{1, X_{j,1}, X_{j,2}\} \end{cases}$$

$$Y_{k,n} = \beta_{k,3} \cdot \alpha_{3,n}$$

формула

$Y_{k,n}$ - известна.

$\beta_{k,3}$ - задаётся из соображений интерпретации.

$\alpha_{3,n}$ - нужно определить.

$$(1, x_{j1}, x_{j2})$$

$(1, 1, 0)$ - для успевающих

$(1, 0, -1)$ - для неуспевающих

$(1, 1, -1)$ - для остальных

$$\begin{pmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,16} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,16} \\ . & . & . & . \\ Y_{461,1} & Y_{461,2} & \dots & Y_{461,16} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \\ . & . & . \\ 1 & 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_1 6 \\ \alpha_{11} & \alpha_{21} & \dots & \alpha_{16,1} \\ \alpha_{12} & \alpha_{22} & \dots & \alpha_{16,2} \end{pmatrix}$$

$$Y_{n,k} = \beta_{n,3} \cdot \alpha_{3,k}$$

$$Y_{k,n}^T \cdot Y_{n,k} = Y_{k,n}^T \cdot \beta_{n,3} \cdot \alpha_{3,k}$$

$$Y_{k,n} = \alpha_{k,3} \cdot \beta_{3,n}?$$

… все фигня и сублимация, саша, давай по новой

$$Y_{n,k} \cdot Y_{k,n}^T = \beta_{n,3} \cdot \alpha_{3,k} \cdot Y_{k,n}^T$$

$$W_{n,n} = g_{n,k} \cdot Y_{k,n}^T$$

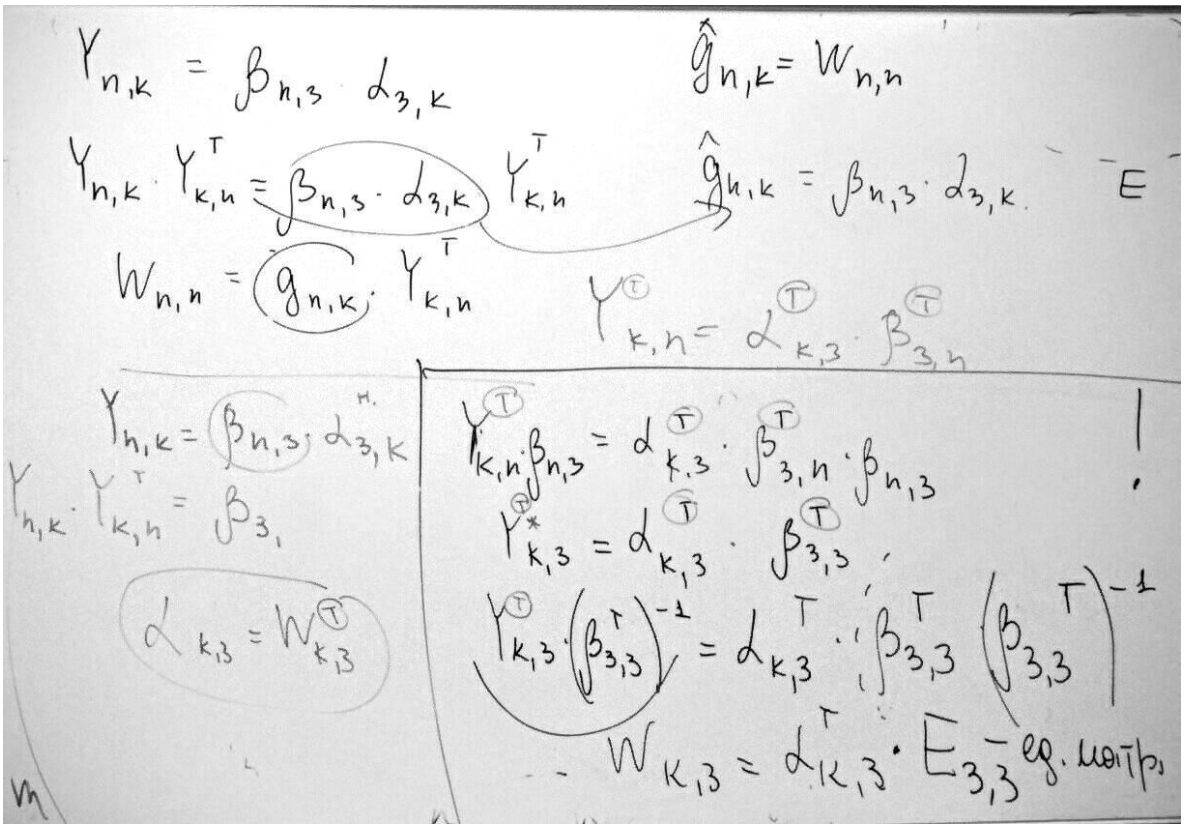
$$\underline{g}_{n,k} = W_{n,n} \cdot (Y^{-1})_{n,k}$$

$$\underline{g}_{n,k} = \beta_{n,3} \cdot \alpha_{3,k}$$

По результатам оценивания для фактора “ночная подготовка к экзаменам” $\mu = 1.5$, $\alpha_1 = 0.2$, $\alpha_2 = 2.6$.

Далее начинается индентификация каждого фактора. Ночное обучение однозначно относится к деструктивным факторам. А часть факторов попадут в нейтральные.

Таким образом, была проведена классификация факторов, выделены оказывающие положительное влияние, негативное и нейтральное. При этом, межвидовая дисперсия факторов заранее определяется выбором веса фактора(или факторов?), а внутривидовая дисперсия разброса определяется разницей между абсолютными значениями $X_{i,j}$ для каждого j-ого фактора.



8. Дисперсионный анализ

ANOVA (ANalysis Of VAriables)

Метод предложен Фишером, согласно этому методу полное изменение анализируемого вектора данных выражается как:

$\underline{y}$: $\sum_{i=1}^{n} \quad (y_i - \underline{y})^2$(*)

и (*) представима в виде суммы связанных с результатом действия различных классификационных факторов.

$\underline{y}$: $y_1, y_2, \ldots, y_n$.

Пусть на результат влияют два фактора: aи b, тогда одномерный вектор $\underline{y}$ представим в виде таблицы $m \times n$ (m экспериментов, n объектов). Сумма (*) представима в виде:

$$\sum_{i=1}^{n} \quad (y_{ij} - y_{..})^2 = \sum_{i=1}^{n} \quad (y_{i.} - y_{..})^2\{A\} + \sum_{i=1}^{n} \quad (y_{.j} - y_{..})^2\{B\} + \sum_{i=1}^{n} \quad (y_{i.} - y_{.j} + y_{..})^2\{AB\}.$$

Элементы с индексом точка рассматриваются как результаты ... y_{ij}по строкам, y_{ji}по столбцам, $y_{..}$ по всей таблице. Если величина $\{AB\}$ оказывается незначительной, говорят об отсутствии взаимодействия факторов A и B. Аналогично, близкое к нулю значение $\{AB\}$ говорит о слабом влиянии фактора A на результат.

Пример

m=20, $\underline{x} = 29.23$, $S_1^2 = 5.62$

n=10, $\underline{y} = 27.56$, $S_2^2 = 2.19$

В этой задаче выборки x и y принадлежат нормальной генеральной совокупности с неизвестными параметрами (μ, σ).

$(m-1)S_1^2 = \sum_{i=1}^{m} (x_i - \underline{x})^2 \approx \sum_{i=1}^{m} (x_i^2 - m\underline{x}^2)$,

$(n-1)S_2^2 = \sum_{i=1}^{n} (y_i - \underline{y})^2 \approx \sum_{i=1}^{n} (y_i^2 - n\underline{y}^2)$.

$\frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$ — результирующая дисперсия.

$$28S^2 = (19 * 5.62) + (9 * 2.19)$$

$$\hat{S}^2 = 4.52$$

$$\{A\} = 2.49$$

$$\{B\} = 1.13$$

$$\{AB\} = 4.52 - 2.49 - 1.13 = 0.9$$

S_1^2 можно рассматривать как внутригрупповую дисперсию выборки 1, S_2^2 соответственно как внутригрупповую дисперсию выборки 2, S^2 как результирующую дисперсию, а разность $S^2 - S_1^2 - S_2^2$ как дисперсию, обусловленную взаимным влиянием факторов.

$$S^2\left(\frac{1}{m-1} + \frac{1}{n-1}\right) = \frac{S^2}{m-1} + \frac{S^2}{n-1}$$

$$\frac{1}{m-1} + \frac{1}{n-1} = \frac{m+n-1}{(m-1)(n-1)}$$

$\frac{m+n-2}{(n-1)(m-1)}(\underline{x} - \underline{y})(**)$

По формуле (**) определяется различие между выборками (межгрупповая дисперсия), при этом внутригрупповая дисперсия определяется как сумма величин $\{A\}$и $\{B\}$.

Если величина (**) незначительна, то говорят о преобладании различий внутри выборок, но не между выборками. Можно рассматривать такие данные как одну группу.

В случае k факторов (k выборок) полная сумма квадратов $\sum_S \sum_r (x_{rs} - x_{..})^2 = \sum_S \sum_r (x_{rs} - x_{.s})^2 + \sum_S n_s(x_{.s} - x_{..})^2$ (Влияние s-ого фактора + взаимное влияние всех факторов).

9. Ранговые коэффициенты корреляции Спирмена и Пирсона

9.1. Коэффициент корреляции (Пирсона).

$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y})}{\sum_{i=1}^{n} \sqrt{(x_i - \underline{x})^2 (y_i - \underline{y})^2}}$, x,y- случайные величины, $0 \leq \rho \leq 1$.

Коэффициент корреляции применяется в случае линейной зависимости между случайными величинами. При нелинейной зависимости - не применяется, даёт неверный результат. Используется в предположении, что выборки x и y - одинакового объёма.

9.2. Ранговый коэффициент корреляции Пирсона.

Пример
Дано:
Данные по защитам дипломных работ в вузе О с 2003 по 2007 год. Указывается процент отличных оценок, хороших, удовлетворительных и общее количество выпускников. Определить, существует ли зависимость между числом выпускников и распределением оценок на защите дипломных проектов (признак квалификации).

$\hat{\chi}^2 = \sum_i \quad \sum_j \quad \frac{(n_{ij} - p)^2}{p}$, $p = \frac{n_{.j} * n_{i.}}{n}$ (*)

	Отл, %	Хор, %	Удовл, %	Общее кол-во
2003	68	25	7	1485
2004	40	40	20	1412
2005	55	33	12	1388
2006	59	28	13	1435
2007	48	37	15	1422

$$D = \sqrt{(1485 - 68)^2 + (1412 - 40)^2 + (1388 - 55)^2 + (1435 - 59)^2 + (1422 - 48)^2}$$

Рассчёт статистики $\hat{\chi}^2$:

Отл 2003 $\frac{(1485 - 1485 * 0{,}68)^2}{1485} \approx 0{,}1$

Допустим сумма первого столбца 3.85, второго - 2.95, третьего - 0.6, общая сумма - 7.4.

В рамках решения задачи проверяется гипотеза о зависимости признаков(нульгипотеза). Альтернативной является гипотеза о независимости признаков. Для проверки гипотезы используется статистика $\hat{\chi}^2$, по таблицам распределений при заданном уровне значимости... и хуйню стала нести.

Вообще хуету несёт, пиздец просто.
Рассчитанное значение статистики находится в основной области, то есть признак квалификации и общее количество выпускников являются зависимыми величинами.
Следует отметить, что количество выпускников за 5 лет менялось мало, так же как и статистика квалификационного признака, в силу чего достоверный результат можно получить при наблюдениях, связанных с изменением общего количества выпускников.
В (*) величина p может вычисляться по-разному, в зависимости от алгоритмов формирования признака квалификации.

9.3. Ранговый коэффициент Спирмена.

Данные по защите дипломного проекта группы ВУЗа У и данные по оценкам абитуриентов(средний балл, полученный на вступительных экзаменах).
ДП - double penetration по дипломному проекту
ВЭ - вступительный экзамен

Ранги	
2	1
3	2
4	3
5	4

<3,25	1
3,25-3,50	2
3,75-4,25	3
4,5-5	4

	ДП, баллы	ДП, ранг	ВЭ, баллы	ВЭ, ранг	Разница^2
1	5	4	4,25	3	1
2	5	4	4,5	4	0
3	4	3	3,25	2	1

4	5	4	4,25	3	1
5	4	3	4,5	4	1
6	4	3	4	3	0
7	3	2	3	1	1
8	4	3	3,5	2	1
9	4	3	3,5	2	1
10	2	1	3,25	2	1
11	4	3	4	3	0
12	4	3	3,5	2	1
13	5	4	4,5	4	0
14	5	4	4,75	4	0
15	3	2	3	1	1
16	5	4	4,25	3	1
17	5	4	4,73	4	0
18	5	4	5	4	0
19	4	3	4	3	0
20	3	2	3,25	2	0

сумма разложения рангов = 10.

$(R_{\text{ДП}} - R_{\text{ВЭ}})^2 = 10$

В рамках задачи проверяется гипотеза независимости признаков (нульгипотеза), альтернативная гипотеза - функция зависимы. Кароче всё равно надо спросить у кого-то..

$$\rho_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - S_i)^2 \approx 0.9$$

Спирмен показал, что величина $\sqrt{n-1}\rho_S$ распределена по стандартному нормальному закону ~N(0,1). Определяем уровень значимости(0,05) по статистической таблице, определяем размер критической области(1,64).

Анализируемая величина порядка 4,15. Внутри отрезка - критическая область, а вне - основная.

Мы отвергаем основную гипотезу в пользу альтернативной. Гипотеза зависимости признаков отвергается.

За счет ранжирования можно выделять группы данных одинакового количества для одной и другой выборки и сравнивать не элементы выборок, а их ранги, что позволяет перейти к задаче сравнения одновременно наблюдаемых параметров.

$$u = f(x_1) - f(x_2, x_3)$$

x1 - число вошедших в течение дня
x2 - число купивших билет(по категориям)
x3 - число прошедших через турникеты

$f(x_2, x_3) = \hat{S}(x_2)$.

$$x_1^{(6)} = kx_3^{(6)} \rightarrow \hat{k} - ?$$
$$x_1 = kx_3$$

$x_1 = \hat{k}x_3$ - по всем станциям (1.599726)

По известному значению $x_3 = \frac{x_1}{\hat{k}} = \frac{1.599726}{\hat{k}} = \frac{1.599726}{1{,}427} = 1{,}121041$

$$x_1 = 1{,}427x_3^{0{,}975}$$

$x_3 \in [21.772091, 41.366973]$, $\gamma = 0.999$

С вероятностью $\gamma = 0.999$ оплаченные поездки(реальный транспортный доход) оцениваются интервалом $[21.772091, 41.366973]$. Реальный дневной доход в 2013 году составил 25 миллионов крон. ... за счёт оценки стоимости долговременных билетов. Доверительная вероятность оценивается по диапазону и будет меньше, чем 0,999. Выявлены зависимости между числом вошедших в течение дня и прошедших через турникет, структура проходящих вне турникета. Получена оценка средней поездки в 2013 году, что позволяет дать рекомендации по снижению убытков на следующий рассчётный период.

Отчёт
Постановка задачи : чешский метрополитен, оценить убыток(A=35млн-21.77 - верхний убыток [0,A]; верхняя граница прибыли B=41.366-35млн [0,B])
дана выборка наблюдаемых и ненаблюдаемых параметров
нужно было установить связь, чтобы посчитать доход
как мы считаем?(модиф метод моментов)
тестировали на своих данных
понятно, да?
ну просто, мы писали, раньше у вас есть
теоретичекие сведения должны быть о модиф методе моментов
рассказать про фиктивный тест и про тест на наших данных

10. Модель факторного анализа

$$y = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n$$
$$(*)\ x_i = \beta_0 + \ldots + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} + \ldots + \beta_n x_n$$

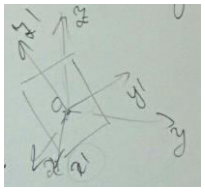
Полагается, что значение каждого признака x_i могут быть выражены (*), как взвешенные суммы других признаков(латентные признаки/переменные/факторы), количество которых может быть меньше, чем число исходных признаков. ε_i (неспецифический фактор) остаточный член ряда, определяющий влияние неучтённых факторов. Можно говорить, что дисперсия x_iзависит от $\sigma_i(\varepsilon_i)$.

$$(*)\ x_i = \beta_0 + \ldots + \beta_{i-1}x_{i-1} + \beta_{i+1}x_{i+1} + \ldots + \beta_k x_k + \varepsilon_i$$

$$(*)\begin{cases} x_i = \beta_0 + \ldots \beta_{i-1}x_{i-1} + \beta_{i+1}x_{i+1} \ldots \beta_k x_k + \varepsilon_i \\ \sigma_i = \sigma(\varepsilon_i) \end{cases}$$

Коэффициенты при переменных называются нагрузкой фактора, переменные x_i— факторами, велечины $\varepsilon_i, i = 1, \ldots, k$ независимы друг от друга и от любоого фактора x_i. Можно наложить условие для n признаков, оптимальное число факторов kопределяется эмпирической формулой: $n - k < \frac{n+k}{2}$ (например, оставим 4 фактора из 10 признаков).

Сумма квадратов нагрузок основной модели факторного анализа называется общностью соответсвующего признака x_i, чем больше это значение, тем лучше описывается решение задачи выделенным фактором. Общность есть часть дисперсии признака, которую объясняют интерпретированные факторы, в свою очередь, ε_iпоказывает какой вклад внесли неитерпретированные факторы. В связи с этим общность называют характеристикой специфичных признаков, в каечтве альтернативы — неспецифичного признака ε_i.

Основное соотношение факторного анализа (*) показывает, что коэффициент корреляции двух любых признаков можно выразить суммой произведения нагрузок некоррелированных факторов.Формально (*) имеет k неизвестных и k уравнений, то есть задача оценки нагрузки решается однозначно, с другой стороны, наложение дополнительных условий в (*) может привести к увеличению неизвестных и неоднозначному решению, в этом случае можно уменьшить количество факторов за счет вращения системы в заданной системе координат. Во избежание снижения точности, осуществляют поворот (рисунок 1)  гиперсферы данных в пространстве при переходе к новой ортогональной системе координат, в этом случае разброс ряда переменных оказывается незначительным или равным нулю, в силу чего измененением этой координаты можно пренебречь, такой подход называется вращением факторов и лежит в основе метода главных компонент. В предположении, что наборы коэффициентов нагрузок β^j для каждого jфактора различны, часто применяют сингулярный анализ, формируют матрицу нагрузок B.

$$B = \begin{bmatrix} \beta_1^1 & \beta_2^1 & \dots & \beta_k^1 \\ . & . & \dots & . \\ \beta_1^k & \beta_2^k & \dots & \beta_k^k \end{bmatrix}$$

$B = USV^T$, U, V — треугольные, S — диагональная. На главной диагонали S расположены сингулярные числа. Обычно, принято располагать на диагонали в порядке убывания.

Признаки, соответствующие минимальному сингулярному числу имеют минимальный разброс, следовательно минимальное влияние на решение задачи — пренебрегаем.

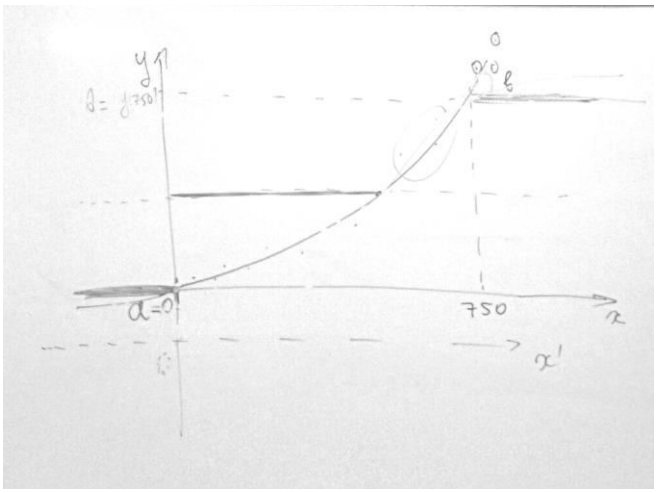
Факторный анализ как процедура реализован во всех пакетах символьных вычеслений, статистической обработки информации, процедуры работают однотипно: начинают с однофакторной модели, затем проверяется насколько корреляционная матрица, восстановленная по однофактроной модели отличается от корреляционной матрицы исходных данных, если однофакторная модель признается неудовлетворительной, то переходят к двуфакторной модели, на каждом этапе проверяется соответсвие корреляционных матриц до тех пор, пока не будет выбрано оптимальное количество параметров, либо достигнуто максимальное количество факторов. Вращение факторов может производится разными способами, что за частую приводит к неподдающимся содержательной интерпретации факторам, после выбора основных факторов, вращение продолжают до тех пор, пока факторы не окажутся поддающимися интерпретации. Можно, например, вращать таким образом, чтобы исчезли трудноинтерпретируемые нагрузки, в ряде случаев исследователи жертвуют условием коррелируемости (взаимосвязи) факторов.

Основным недостатком метода является неоднозначность полученного решения, в ряде предметных областей факторный анализ — привычный и единственных инструментарий исследования. Для выбора решения проводят анализ вклада пары наиболее значимых факторов методом дисперсионного анализа, если дисперсионный и факторный анализы дают сходные факторы, то решение можно считать однозначным.

11. Цензурирование и анализ выбросов

Цензурирование выборки — замена реальных значений случайных величин выбранными аналогами.

Анализ выбросов — изъятие значений случайных величин из результатов эксперимента на основе статистического анализа данных.



В случае, если преполагаемые реальные значения близки к заданому значению левой границы выборки a, но не могут быть измерены, реальные значения заменяются на величину a(цензурирование слева). Если предполагаемые значения близки к правой заданной границе выборки b, но не могут быть измерены, то они заменяются на значения b(цензурирование справа). Возможна ситуация, когда выборке требуется цензурирование слева и справа.

Исследователь может руководствоваться сведениями о виде функции, при этом, как правило, под цензурирование и слева, с права попадает не более 10% процентов исследуемых значений, в противном случае рекомендуется изменить условия проведения эксперимента.

Анализ выбросов является процедурой альтернативной цензурированию и позволяет выявить точки, искажающие решения задач.

Существует ряд подходов к анализу выбросов, которые можно условно классифицировать:

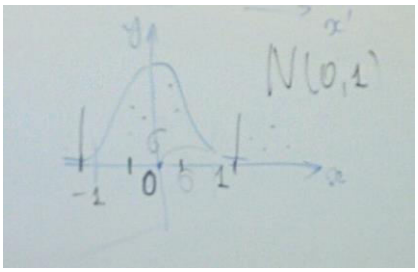
A. базирующиеся на виде закона распределения;
B. базирующиеся на построении доверительного интервала.

К группе А относятся:

1. Метод, основанный на предварительном знании вида функции распределения случайной величины, наблюдаемой в рамках эксперимента;

В курсе лекций был рассмотрен пример о проверке равномерности выборки генеральной совокупности, распределенной нормально с выбранными параметрами μ и σ.

2. Если закон распределения выборки $x \in X: N_X(\mu, \sigma)$, то анализ выбросов производится на основе правила трёх σ. Процедур, относящихся к классу А немного, так как они специфичны для конкретного закона распределения. Кроме того закон распределения может быть неизвестен.



К группе В относятся:

1. Альтернативой, обеспечивающей анализ выбросов является подход, связанный с построением доверительных интервалов. Если значительная часть элементов выборки оказывается вне интервала, то говорят о наличии выбросов. $[\underline{x} - k\hat{S}, \underline{x} + k\hat{S}], \gamma = 0.99, k = 1.55 + 0.8lg(\frac{N}{10})\sqrt{e-1}(1)$
2. Существует дост много вариантов оценок коэффициента k, в ряде случаев анализируют выборочную функцию распределения и фиксируют квантили заданного уровня

Если при проведении другого эксперимента значимо большее количество элементов выборки, чем в контрольной, оказывается вне диапазона $[K(\alpha_1), K(\alpha_2)]$, то говорят о наличии выбросов.

Недостатком методов группы В является выявление значительного количества выбросов в выборке.

12. Моделирование на ЭВМ случайных величин, векторов, процессов

В ряде случае возникает необходимость моделирование данных на ЭВМ, в этом случае говорят о применении генератора псевдослучайных чисел по заданному закону распределения.

Различают 3 вида генераторов случайных чисел:

1. физические (бросание монеты, физические приборы и датчики, таймеры);
   достоинства: точность моделирования;
   недостатки: ограниченный ряд законов распределения, неудобство использования;
2. табличные;
   достоинства: высокая точность моделирования данных;
   недостатки: большой объём справочной информации, которую необходимо хранить и считывать из файла;
3. математические (ГПСЧ);

Первый (арифметический) генератор случайных чисел был предложен фон Нейманом, имел очень малый период (быстро возникало зацикливание): либо возникали нули, либо число переходило само в себя.

Модификации арифметического способа предлагали осуществление сдвига влево или вправо, что позволяло увеличить период метода (до тысячи значений без зацикливания),

тем не менее, подход был малопригоден для решения задач на компьютере. В настоящее время используется линейный конгруэнтный подход.
$X_{n+1} = (aX_n + b)\ mod\ m$ (mod — операция взятия по модулю m, даёт остаток от деления результата $aX_n + b$ на m).
Предложил подход (1949 год), оптимальный вариант получили в 90-е годы. Предложенный генератор предполагает выбор 3 чисел a, b, m, начальное значение X_0 берётся с таймера. При удачном выборе начальных чисел, генерируемая последовательность содержит независимые случайные величины.
$a = 134775813, b = 1, m = 2^{32}$.
Следует отметить, что начальное приближение формально может быть выбрано пользователем. Если в двух разных точках последовательности получается одно и то же значение, то далее последовательности формируются одинаково. В этом случае лучше начальное приближение менять случайным образом.

Пример:
Генератор Парка-Милера (с недостатком переполнения разрядной сетки).
$b = 0, m = 2^{31} - 1, a\ = 16807$.

Предложенные математические генераторы дают выборку с равномерным законом распределения.
Базовые законы распределения, как правило, строят на основе нескольких выборок равномерно распределенных величин или способом обращений. Допустим, необходимо сформировать выборку из нормально распределенных случайных величин с заданными параметрами.

1. Метод на основе центральной предельной теореме (ЦПТ). Известно, что сумма нескольких независимых случайных величин, равномерно распределенных в интервале $[0,1]$ асимптотически стремится к нормальному распределению, то есть имеет асимптотически нормальное распределение.
   $r_i: R[0,1]$
   $x = \sum_{i=1}^{n} \quad r_i, x \in N(\mu, \sigma)$
   $[2,1,5,3,4,6]$
   $$\frac{1}{6}(1.5^2 + 2.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 2.5^2) = \frac{1}{6}(4.5 + 0.5 + 12.5) \approx 2.91 \approx 1.7$$
   $$\mu = 3.5, \sigma = 1.7$$
2. Метод обращений $F(x) \sim f(x)$
   $(*)\ F(x) = 1 - e^{-\lambda x}$
   $f(x) = \lambda e^{-\lambda x}$
   $$\lambda = \frac{1}{x}$$
   Рассчитывается выборочное среднее и строится выборочная функция распределения, которая может быть описана аналитической функцией вида (*). Строится функция, обратная к функции распределения $F^{-1}(x)$, которая выражается через функцию плотности распределения с неизвестным параметром λ, откуда может быть определён параметр λ.

$$F^{-1}(x) \Rightarrow x = 1 - F(x) = e^{-\lambda x} = \frac{f(x)}{\lambda}$$
$$F^{-1}(x) = f(x) = \frac{1}{\lambda} f'(x) \to 0 + \frac{f(x)}{f'(x)} = \frac{1}{\lambda}$$

на основе предположения о чем там про 0 и уравнени чего-там касательно рассматриваемой как сумма приращений аргумента x

Метод обращений реализуем для небольшого круга функций и, как правило, используется для экспоненциального закона.

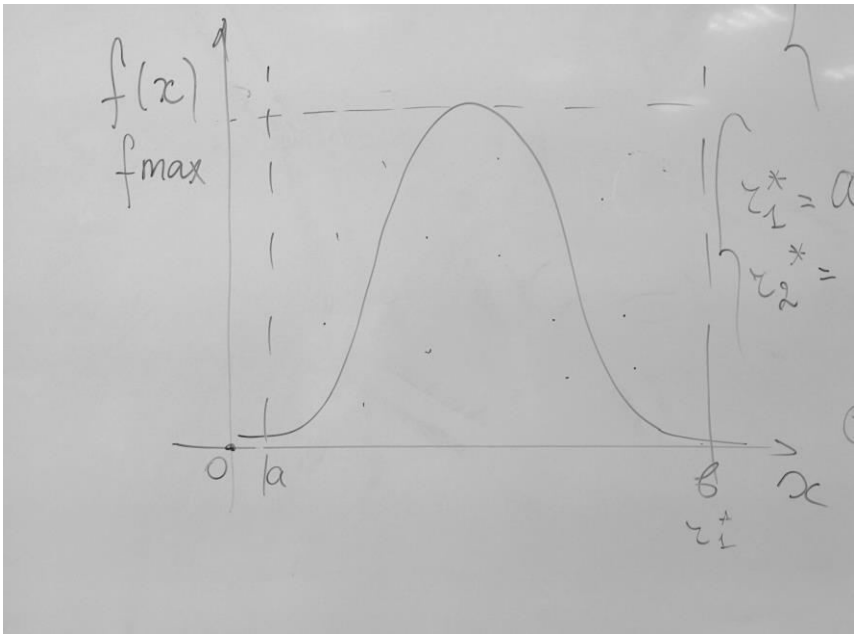
3. Метод Неймана. Подход является универсальным для произвольной функции распределения.

$r_1 : R[0,1] \to r_1^* : R[a,b]$

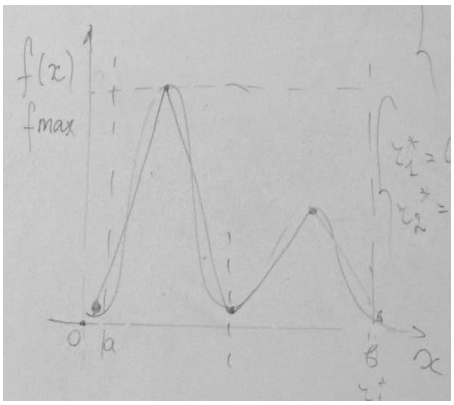
$r_2 : R[0,1] \to r_2^1 : R[0, f_{max}]$

$$r_1^* = a = (b-a) \cdot r_1$$
$$r_2^* = f_{max} \cdot r_2$$

(*)
$$\begin{cases} r_2^* \le f(r_1^*) \\ r_1^* = a + (b-a) \cdot r_1 \\ r_2^* = f_{max} \cdot r_2 \end{cases}$$



Если пары точек удовлетворяют условию ограничения системы (*), то соответсветсвующие им значения r_1^* составляют выборку из генеральной совокупности, распределенной по заданому закону, в противном случае — отбрасываются.



Если распределение имеет сложный вид, его докомпозируют на простые по форме составляющие, интерполируют простыми распределением, чаще равномерным. Такой метод называется методом суперпозиции.

В предположении, что случайный вектор представляет собой набор зависмых случайных величин с заданным законом распределения, моделирование случайного вектора не отличается от подхода к моделированию случайных процессов, т.к. случайные процессы (функции) на ЭВМ дискретизируются. Рассмотрим моделирование случайного процесса с показательным распределением, в этом случае получают последовательность n независимых случайных величин (с нормальным стандартным распределением), после чего строится следующая реккурентная зависимость:

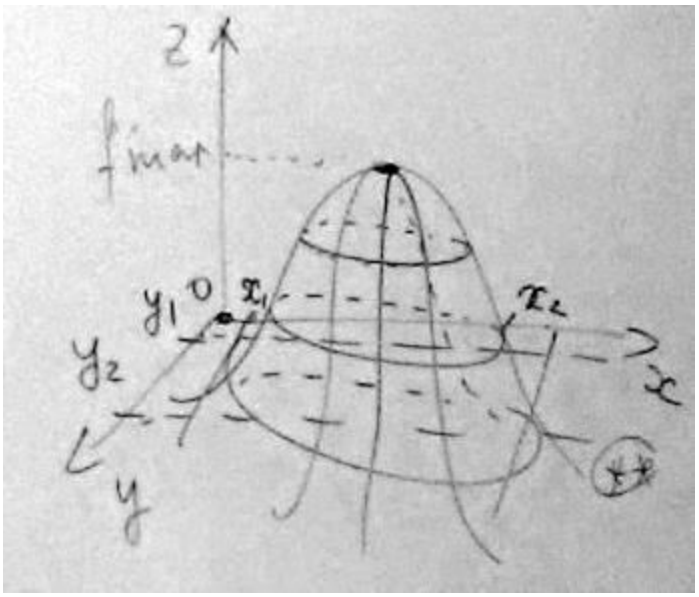
$$\xi(t_n) = \rho_n \xi(t_{n-1}) + \sqrt{1-\rho_n^2}\, x[n],$$

$$\rho_n = e^{-\lambda(t_n - t_{n-1})},$$

$$\lambda = \frac{\pi}{2\Delta t}.$$

Показательное распределение (экспоненциальный закон) является примером моделирования стационарного случайного процесса с заданным законом распределения. Другие законы распределения предполагают более сложную систему описания, в силу чего в ряде случаев в задаче моделируются случайные процессы в иной классификации.

Универсальным подходом к моделированию стационарного случайного процесса с заданными законом распредления является обощение метода фон Неймана на n-мерный вектор. Формируется набор из $n+1$ случайной величины, распределённый равномерно на интервале $[0,1]$.



Проводится преобразование: $r_1 \to r_1^*, \ldots, r_n \to r_n^*$, где значения со звёздой изменяются в любом заданном диапазоне.

(**) $r_{n+1}^* \le f(r_1^*, \ldots, r_n^*)$

Если выполняется условие (**), то набор сохраняется и координата r_{n+1} считается распределённой по определённому закону в диапазоне $[0,1]$. В противном случае набор отбрасывается. Недостатками метода являются усечение функции плотности распределения и получение набора независимых случайных величин.

Другая классификация (марковские/немарковские процессы) предполагает моделирование без учёта закона распределения набора случайных величин.

Марковский случайный процесс — это случайный процесс, реализация которого в заданный момент времени известна. Переход процесса в новое состояние при известном значении текущей реализации не зависит от его прошлых состояний. Различают понятие марковской цепи (частный случай марковского процесса), когда пространство состояний дискретно.

Кроме того, марковский процесс рассматривают как авторегрессию первого порядка.

Определение марковского процесса по Венцель: “будущее” процесса зависит от “прошлого” только через его “настоящее”.

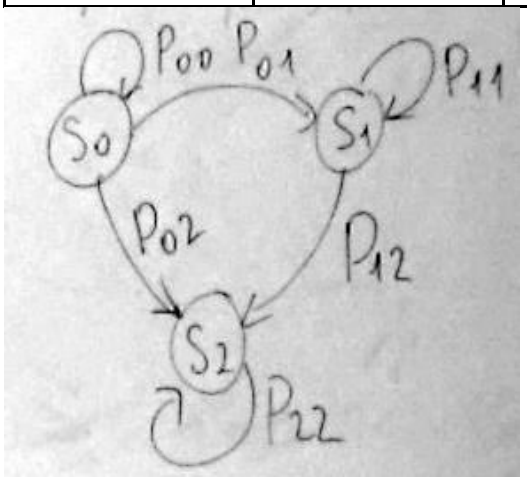
$$x_{n+1} = ax_n + b$$

Пример.

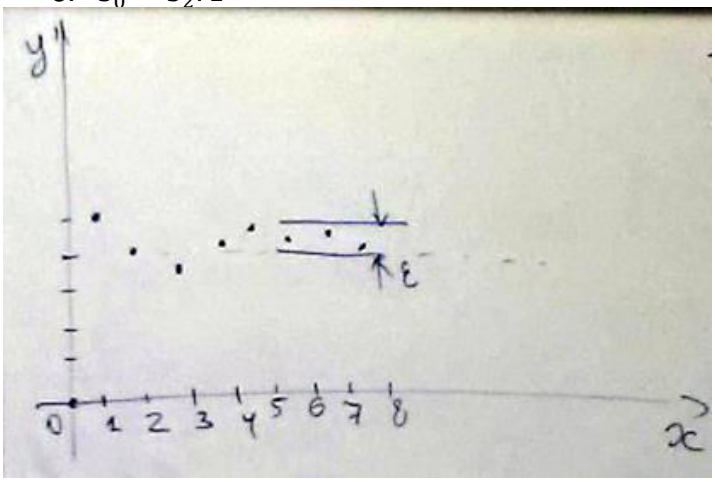
Известны вероятности перехода из одного состояния в другое(состояния: S_0 — новая мишень, S_1 — повреждённая мишень, S_2 — поражённая мишень). Необходимо определить $\underline{k}$ — среднее количество снарядов, необходимых для поражения цели.

	S_0	S_1	S_2	

S_0	P_{00}	P_{01}	P_{02}	$\sum_{i=1}^{3} P_{0i} = 1$
S_1	0	P_{11}	P_{12}	$\sum_{i=1}^{3} P_{1i} = 1$
S_2	0	0	1	$\sum_{i=1}^{3} P_{2i} = 1$



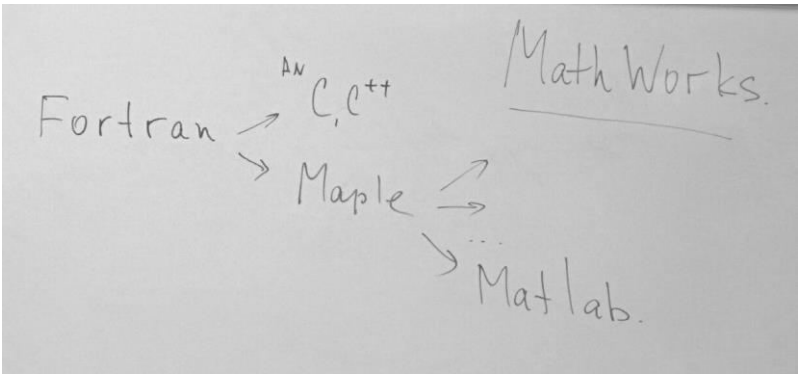
1. $S_0 \to S_0 \to S_1 \to S_1 \to S_1 \to S_2 : 5$
2. $S_0 \to S_0 \to S_0 \to S_2 : 3$
3. $S_0 \to S_1 \to S_1 \to S_2 : 3$
4. $S_0 \to S_0 \to S_1 \to S_1 \to S_1 \to S_2 : 5$
5. $S_0 \to S_0 \to S_0 \to S_1 \to S_1 \to S_1 \to S_2 : 6$
6. $S_0 \to S_0 \to S_1 \to S_1 \to S_2 : 4$
7. $S_0 \to S_0 \to S_0 \to S_0 \to S_1 \to S_1 \to S_2 : 6$
8. $S_0 \to S_2 : 1$



При увеличении числа реализаций случайного процесса оцениваемая величина стремится к величине $a : k \to a$, и при бесконечном числе реализаций — подстигает его.

Для остановки счёта на ЦВМ пользуются следующим правилом: если 3 последние точки попадают в заданный диапазон с разницей границ не превышающей ε, то счёт завершается, а оценкой $\hat{a}$ является середина диапазона.
В данном случае величина равняется 4.25, выбранный диапазон $[4; 4.5]$, так как речь идёт о ресурсах, то величина округляется в большую сторону.
Ответ: в среднем потребуется 5 снарядов для поражения цели.

13. Классификация современных средств моделирования на примере пакета MathWorks



Решения компании MathWorks представляют собой современный инструментарий проектирования и имитации сложных систем на основе заданной или построенной математической модели. Позволяют создавать сложные многодоменные проекты, включающие приложения для обработки изображений и сигналов, математических расчётов, расчёта технической оснастки проекта, обмена данными и т.д. MathWorks предполагает описание всех аппаратных программных и алгоритмических средств, участвующих в моделировании как объектов, характеризуя их наборами данных (а также моделями данных), функциями и возникающими в процессе функционирования системы событиями. Такой подход называется объектно-ориентированным проектированием.

13.1. Построение модели сложных систем

1. Наследуя принципы среды моделирования Simulink, основанные на графических блок-схемах, MathWorks позволяет выбрать структуру описываемой системы и установить связи (разного типа, как в UML) между компонентами системы.
2. Вводится набор дескрипторов, определяющие каждую компоненту как объект. То есть каждый блок компоненты может быть описан физически, математическими объектами, дискретными событиями или графиками состояния.
3. Оптимизация схемы — удаление/объединение блоков, функционирование которых не влияет на решение задачи.
4. Выбирается способ имитации системы, реализуются алгоритмы посредством библиотеки алгоритмов, наследованной от MATLAB и Simulink.

Достоинства пакета MathWorks:

1. интеграция большого объема вычислительных методов;
2. удобный графический интерфейс;
3. интерфейс с другими языками программирования, в том числе импорт/экспорт программ из/в C, C++;

4. возможна автоматическая генерация на C, C++, код генерируется непосредственно из модели сложной системы, что удобно для развёртывания и прототипирования;
5. встроенный язык для распараллеливания вычислений, в том числе распараллеливание реально поступающих данных с датчиков;
6. встроенная система тестирования.

Недостатки пакета MathWorks:

1. низкая точность вычислений;
2. значительное количество ошибок возникает при распараллеливании вычислений и данных;
3. ограничен интерфейс с иными средствами программирования.

Объектно-ориентированное моделирование позволяет в короткий период времени создать прототип системы, нормализовать его, оптимизировать, а также, верифицировать, обеспечить валидацию требований и осуществлить тестирование модели.

Целью процесса является выявление ошибок на ранних этапах моделирования. Для каждой компоненты системы вводится так называемая спецификация, включающая помимо атрибутов и операций наборы входных\выходных данных в заданный или формируемый момент времени. Цензура ебанная. Верификация выполняется на всех этапах функционирования системы для каждой компоненты. Верификация дожна учитыват взаимное влияние компонент и изменения условий моделирования. Процесс довольно трудоёмкий и может быть унифицирован за счёт наложения однотипных условий на однотипные компоненты (и процессы).

К проблемам тестирования модели относятся неверно сформулированные требования. В том числе и противоречивые требования. Возникает необходимость в ранней валидации требований. Как правило, модель содержит ограничения, накладываемые на переменные с учетом физической природы, изменений окружающей среды или изменений природы, а также накопление вычислительной погрешности при реализации модели.

В случае, если эти составляющие на практике учесть представляется не возможным или сложно, то исследователь вводит доверительный интервал с заданным коэффициентом доверия для исследуемой величны. В ряде случаев исследователь затрудняется в оценке границ доверительного интервала с высоким коэффициентом доверия — в этом случае, необходимо построить систему тестов для контроля за исследуемыми величинами. В первую очередь создаются тесты системного уровня, обеспечивающие тестирование модели в соответствии с системными требованиями, генерируется пространство случайных параметров, при реализации модели величины, исследуемые на таких пространствах должны попадать в заданный диапазон. В MathWorks реализован модуль System Test, который на основе продукта Simulink Verification & Validation, разработчик может связать свою схему модели и диапазоны изменений величин со стандартными процедурами и тестами этих продуктов

14. Математические основы теории массового обслуживания

Создатетем теории массового обслуживания считается советский математик А.Я. Хинчин. Стартом для его исследований явились работы математиков английской анктуарной школы (соц. страхование и аналогичные риски... процессы гибели и размножения) и копенгагенсткой телефонной компании (начало 20 века) (Ерланг - глава отдела).

Пусть имеется телефонный узел (устройство\прибор в терминологии СМО), на котором телефонистки соединяют пары телефонных абонентов. При небольшом количестве звонков соединение не требует ожидания. При интенсивном увеличении говорят о СМО с ожиданием. Ожидающее удовлетворение заявки (транзакты в СМО) помещаются в очередь. Очередь может быть ограничена (значение - М заявок). В этом случае, говорят о возможности потери L-заявок. Если считать транзакты равноправными(актуальными являются только моменты поступления заявок), то поток заявок считается однородным.

Если поток однороден и после их обработки дисциплина функционирования системы не меняется, то говорят о потоке без последействия.
$N = \lambda T$ — формула Литтла. (*)
N- среднее количество заявок в системе
λ - интенсивность
T - время обработки
Поток без последействия имеет место если количество обработанных заявок в любом временном интервале $[t, t + \Delta t]$ остаётся постоянным в любом совпадающем по длительности $[t^*, t^* + \Delta t]$ непересекающимся с исходным интервале времени. Поток заявок стационарен, если вероятность обработки n заявок в интервале $[t, t + \Delta t]$ не зависит от времени t, а зависит только от длительности Δt интервала. Однородный стационарный поток без последействия называется простейшим потоком Пуассона. Число событий такого потока распределено по закону Пуассона. ПедоВики.

Мгновенная плотность потока равна пределу отношения среднего числа заявок обработанные в элементарный интервал времени $[t, t + \Delta t]$ к длине этого интервала Δt стремящимся к нулю. В технических приложениях называется интенсивоностью потока.
$\lambda = \frac{M(t)}{\Delta t}$- для простейшего потока.

Формула Литтла позволяет оценить среднее количество заявок в системе. Основными элементами СМО (помимо входного потока\очереди заявок) являются каналы (несколько однотипных приборов обслуживания), выходной поток, фаза обслуживания. В связи с чем системы деляться на одноканальные\многоканальные, на системы с очередями с ожиданием и отказами, по типу равноценности заявок на системы с приоритетом и без (бывают комбинированного типа), по фазам обслуживания на моно- и много-фазаным обслуживанием. На определённых фазах возможно повторное обслуживание одной и той же заявки. По взаимосвязи с потоками заявок системы делятся на открытые (разомкнутые) и закрытые (замкнутые). Если интенсивность входного потока заявок не зависит ни от количества заявок в СМО, ни от количества уже обслуживанных заявок ,то говорят об открытой СМО. Системы, сочетающие в себе свойства многоканальности, многофазности, разомкнутости, классифицируются как сеть массового обслуживания.

Моделируемые системы должны быть эффективными, в связи с чем вводятся показатели

эффективности СМО:

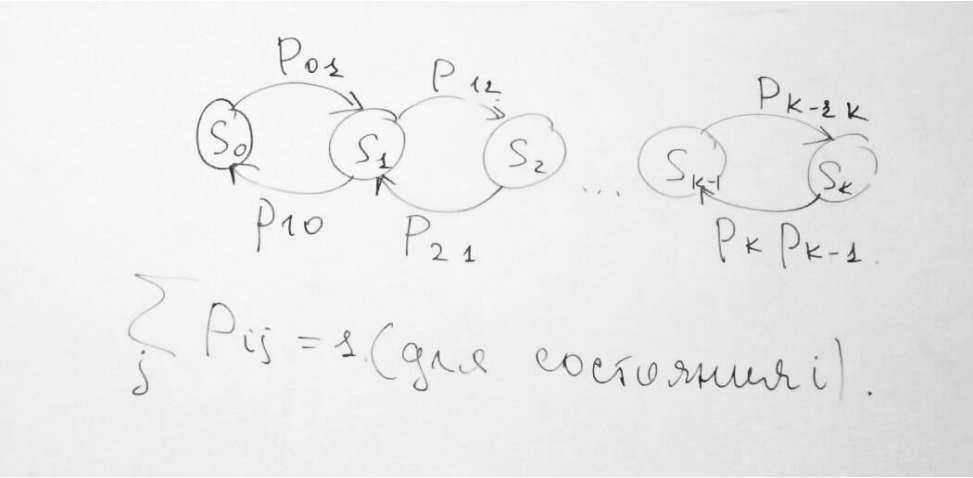
1. абсолютная пропускная способность СМО (среднее количество заявок, обслуживаемых системой в единицу времени);
2. относительная пропускная способность СМО (отношение среднего числа заявок, обслуживаемых в системе в единицу времени, к среднему числу всех заявок, поступивших за это время в СМО);
3. среднее число занятых каналов и коэффициенты их занятости (отношение числа занятых каналов к общему числу каналов, формализуются как показатели эффективности СМО);
4. среднее число свободных каналов и коэффициент простоя (формализуются по показателям занятости каналов);
5. среднее время нахождения заявки в очереди;
6. среднее время нахождения заявки в СМО;
7. дисперсия числа заявок в очереди и в целом в СМО.

Указанные параметры легко формализуются и являются аналитическими моделями СМО.

14.1. Уравнение Колмогорова

Состояние СМО определяется количеством занятых каналов обслуживания и числом мест в очереди. Очевидно, эти параметры целочисленны и меняются дискретно. Хотя изменения происходят в случайные моменты времени. Если система может быть описана марковским процессом, то исследование такой системы существенно упрощается. Вероятность достижения нового состояния не зависит от предыстории процесса. Если можно предположить, что макровская цепь дискретна, то есть переход из состояния в состояние происходит в фиксированные промежутки времени с равными интервалами, то можно рассчитать или прогнозировать вероятность перехода из одного состояния в другое. Если интервал перехода Δt достаточно мал, вероятности для марковских цепей могут быть рассчитаны посредством разностных схем, а в случае непрерывных марковских цепей — дифференциальными уравнениями (ДУ Колмогорова).

Марковский процесс с дискретными состояниями называют процессом гибели и размножения, если он имеет размеченный граф состояний вида:



$\sum_j \quad p_{ij} = 1$(для состояния i)

$$S_0 \to S_1$$

Предполагают, что вероятности $p_{ij}, j = 1,\dots,k, i = 1,\dots,k$

$$\begin{cases} \lambda_{01}p_{01} = \lambda_{10}p_{10} \\ \cancel{\lambda_{01}p_{01}} + \lambda_{12}p_{12} = \cancel{\lambda_{10}p_{10}} + \lambda_{21}p_{21} \\ \cancel{\lambda_{01}p_{01}} + \cancel{\lambda_{12}p_{12}} + \lambda_{23}[_{32} = \cancel{\lambda_{10}p_{10}} + \cancel{\lambda_{21}p_{21}} + \lambda_{32}p_{32} \\ \dots \end{cases}$$

$$(*)\begin{cases} \lambda_{01}p_{01} = \lambda_{10}p_{10} \\ \lambda_{12}p_{12} = \lambda_{21}p_{21} \\ \lambda_{23}p_{23} = \lambda_{32}p_{32} \\ \dots \\ \lambda_{k-1,k}p_{k-1,k} = \lambda_{k,k-1}p_{k,k-1} \end{cases}$$

$$\lambda_{01}p_0 = \lambda_{10}p_0 \Rightarrow \lambda_{01} = \lambda_{10}$$

В (*) $p_{ij} \neq p_{ji}$для большинства практических приложений, в силу чего система имеет k уравнений $2k$ неизвестных.

Для сокращения числа неизвестных в 2 раза равновероятными предполагают переходы из одно и того же состояния в предыдущее и будущее.

$p_{k-1,k} = p_k, p_{k,k-1} = p_{k+1}$

Равновероятным является переход в новое состояние на один шаг и возврат в предыдущее.

$$(**)\begin{cases} \lambda_{01}p_1 = \lambda_{10}p_2 \\ \lambda_{12}p_2 = \lambda_{21}p_3 \\ \lambda_{23}p_3 = \lambda_{32}p4 \\ \dots \\ \lambda_{k-1,k}p_k = \lambda_{k-1,k-2}p_1 \quad , p_{k+1} = p_1 \end{cases}$$

Система (**) называется системой уравнений гибели-размножения.

$$\lambda_{21}p_3 = \frac{\lambda_{21}\lambda_{01}}{\lambda_{10}}p_1 \Rightarrow p_3 = \frac{\lambda_{12}\lambda_{01}}{\lambda_{10}\lambda_{21}}p_1$$

$$p_k = \frac{\lambda_{k-2,k-1}\lambda_{k-3,k-2}\cdot\dots\cdot\lambda_{01}}{\lambda_{k-2,k-3}\cdot\dots\cdot\lambda_{10}} \cdot p_1$$

Все вероятности выражаются через вероятность p_1. Из последнего уравнения (**) выразим p_1:

$$p_1 = \frac{1}{\frac{\lambda_{01}}{\lambda_{10}} + \frac{\lambda_{12}\lambda_{01}}{\lambda_{10}\lambda_{21}} + \dots + \frac{\lambda_{k-2,k-1}\lambda_{k-3,k-2}\cdot\dots\cdot\lambda_{01}}{\lambda_{k-2,k-3}\cdot\dots\cdot\lambda_{10}}}$$

$$(***)\begin{cases} p_1 = \frac{1}{\frac{\lambda_{01}}{\lambda_{10}}+\frac{\lambda_{12}\lambda_{01}}{\lambda_{10}\lambda_{21}}+\ldots+\frac{\lambda_{k-2,k-1}\lambda_{k-3,k-2}\cdot\ldots\cdot\lambda_{01}}{\lambda_{k-2,k-3}\cdot\ldots\cdot\lambda_{10}}} \\ p_2 = \frac{\lambda_{01}}{\lambda_{10}}p_1 \\ p_3 = \frac{\lambda_{12}\lambda_{01}}{\lambda_{10}\lambda_{21}}p_1 \\ p_k = \frac{\lambda_{k-2,k-1}\lambda_{k-3,k-2}\cdot\ldots\cdot\lambda_{01}}{\lambda_{k-2,k-3}\cdot\ldots\cdot\lambda_{10}}\cdot p_1 \end{cases}$$

Предположения, наложенные на модель относительно вероятности перехода, оказываются достаточно жёсткими. И в целом, уравнения системы (***) применительно к практической задаче могут модифицироваться.

В общем случае уравнения для переходной фурнкции марковского случайного процесса описываются как дифференциальные уравнения Колмогорова в преположении, что функция перехода из состояния i в состояния jзависит от моментов переходов s и t $(p_{ij}(s,t),\ t>s,\ t\to s)$.

Колмогоров предложил описывать все процессы в качестве марковских цепей в 1938 году.

1) $\frac{\partial p_{ij}(s,t)}{\partial s} = \sum_k \quad \alpha_{jk}(s)\cdot p_{kj}(s,t)$

2) $\frac{\partial p_{ij}(s,t)}{\partial t} = \sum_k \quad p_{ik}(s,t)\alpha_{kj}(t)$

3) $\alpha_{ij}(s) = \lim_{t\to s}\frac{[p_{ij}(s,t)_{(1-\delta_{ij})}]}{t-s}, t>s$

δ_{ij}-символ Кронекера

Можно сформулировать правила составления уравнений Колмогорова по размеченному графу состояний непрерывной Марковской цепи:

1. Число уравнений в системе равно числу вершин графа ($i=1,\ldots,k, j=1,\ldots,k$), при этом вероятность состояния p_i соответсвует как переходу в последующее, так и возврату в предыдущее состояние.
2. Система дифференциальных уравнений имеет форму Коши.
3. Число слагаемых в правой части равно числу дуг в графе, интерпретирующих переход из i-го состояния в любое другое (кроме самого себя).
4. Переход в новое состояние соответствует слагаемому со знаком “+”, возврат в предыдущее состояние — слагаемое со знаком “-”.
5. Каждое слагаемое представляет собой произведение вероятности i-го состояния плотности $\alpha_{ij}(s)$ и вероятности перехода по данной дуге.
6. Начальные условия для постановки задачи Коши определяются непосредственно начальным состоянием системы, например, в цепочке из состояний $S_0, S_1, \ldots, S_k$ старт начинается из состояния S_2: $p_0(0)=0, p_1(0)=0, p_2(0)=1, \ldots, p_k(0)=0$.