Математическое Моделирование

ИУ9-111

2015

1 Правила форматирования проекта в LaTeX.

Правила оформления темы:

1. Названия в блоках `\section, \subsection, \subsection` должны начинаться с заглавной буквы и заканчиваться симовлом «.».

Правила оформления содержания лекций:

1. Цензура на употребление ненорм. лексику.
2. Каждое предложение должно начинаться **с новой строки**;
3. Вставка тематических цитат должна производиться с помощью блока `\footnote{}` ;
4. Для вставки абзаца, необходимо использовать **пустую строку** (вместо «\\»);
5. Формулы, которые описываются вне текста, должны быть заключены в `\gather` блок (не использовать секцию «`$$ ... $$`»);
6. При использовании списка, каждый итем должен заканчиваться символом «`;`», а последний — «`.`»
7. Для тире использовать «`---`»;
8. Определения понятий должны быть заключены в блок «`{\it ...}`»;
9. Если отсутстсвуют ссылки на рисунок, а также неизвестна подпись — то удаляем теги `\caption, \label`.

Содержание

1 Правила форматирования проекта в LaTeX. 1

2 Введение 4

2.1 Основные задачи мат. статистики 4

3 Точечные оценки параметров 7

3.1 Свойство оценок 7

3.2 Методы построения точечных оценок 9

3.2.1 Метод моментов 10

3.2.2 Метод максимального правдоподобия 13

4 Доверительные интервалы 15

4.1 Методы построения доверительных интервалов 15

5 Лабораторная работа 1 (Разбор) 18

6 Основные понятия проверки статистических гипотез 21

6.1 Критерей Колмагорова-Смирнова 23

7 Методы многомерного статистического анализа (MCA). 25

8 Метрики $\rho(x', x'')$. 27

8.1 Оценка качества кластеризации. 28

8.2 Методы кластерного анализа. 29

8.2.1 Метод K-средних (K-means) 29

8.2.2 Метод Варда 31

8.2.3 Метод ближайшего соседа. 33

8.2.4 Метод наиболее удалённого соседа. 33

8.2.5 Выводы. 34

8.2.6 Лаборторная 34

9 Многомерный дисперсионный анализ 36

10 ANOVA – дисперсионный анализ 41

11 Ранговые коэффициенты корреляции Спирмана и Пирсона. 44

11.1 Коэффициент корреляции Пирсона 44

11.2 Ранговый коэффициент корреляции Пирсона 44

11.3 Ранговый коэффициент Спирмана. 46

12 Факторный анализ. 49

12.1 Модель факторного анализа. 49

13 Цензурирование выборок, и анализ выбросов. 52

14 Моделирование на ЭВМ случайных величин, векторов, процессов. 55

15 Классификация современных средств моделирования на примере пакета MathWorks. 63

15.1 Построение моделей сложных систем. 63

16 Объектно ориентированное моделирование. 65

16.1 Верификация и валидация тестирования. 65

17 Математические основы теории массового обслуживания (ТМО). 67

17.1 Уравнение Колмогорова. 69

18 Процессы гибели-размножения. 71

19 Подготовка к экзамену 74

19.1 Вопрос 74

2 Введение

Изучение математических моделей случайных явлений, или экспериментов, в первую очередь занимаются такие науки, как **мат. статистика**, и **теория Вероятностей**. Задачей мат. статистики является обратными задачами к задачам теории вероятности.

В Теории вероятности (ТВ), после задания того, или иного случайного явления, требуется рассчитать вероятностные характеристики в рамках данной модели. Моделирование проводится на основе результатов эксперимента, называемых **статистическими данными**. В ряде случаев, по результатам эксперимента, требуется лишь уточнить, или модифицировать имеющуюся модель.

В задачах мат. статистики, вероятность того, или иного события известна, и необходимо оценить **параметры эксперимента** (параметры закона распределения случайной величины, в более широком смысле – функцию плотности распределения с.в., и т.п.). Как правило, рассматриваю 3 задачи унификации. $\hat{p} = \frac{m}{n}$

2.1 Основные задачи мат. статистики

- Задача оценки неизвестных параметров по результатам экспериментов.

  Как правило, нужно найти ф-цию от рез-тов эксперимента, зн-е которой является достаточно хорошей оценкой неизвестного, истинного значения параметра. (a – оценка, $\hat{a}$ – оценка параметра)

- Задача интервального оценивания.

  Требуется построить интервал с границами $[a_- \leq a \leq a^-]$ таким образом, чтобы он накрывал неизвестное истинное значение параметра, с заранее заданной вероятностью γ

- Задачи проверки стат. гипотез.

  Требуется, на основе мат. экспериментов, проверить то, или иное предположение относительно вида, и пар-ра ф-ции распределения с.в., и ф-ции пл-ти распределения с.в.

В мат. статистике используется выборочная технология, основанная на *урновой схеме*. Пусть имеется урна, содержащая N чисел (1) $\{X_1, X_2 \dots X_N\}$, называемое *генеральной*

совокупностью объема N. Набор (1) может иметь бесконечную размерность.

Из генеральной совокупности выбирается набор (2) $\{x_1, \ldots, x_2\}, n \leq N$, и называется *выборкой* из генеральной совокупности (1).

Выборка может производиться с возвращением, и без возвращения. Если выборка производится с возвращением, то случайные величины в ней независимы. В противном случае – зависимы. С возвращением тождественного равенства случайная... Терминология сохраняется и в случае бесконечной генеральной совокупности.

Числа выборки (2) располагают обычно в порядке убывания или возрастания, и получаем набор (3) $\{x^{(1)}, \ldots, x^{(n)}\}$, который называется *вариационным рядом*. Чаще всего, в практических задачах анализируется именно вариационный ряд.

Имперической ф-ции распределения, построенной на основе выборки (3), называется ф-ция:

$$\hat{F}(x) = \frac{r(x)}{n}$$

n – общее число элементов выборки,

$r(x)$ – количество эл-тов $x_i : x_i \leq x$

Рис. 1 – Вид имперической функции распределения

Для моделирования требуется теоретическая ф-ция распределения с.в X, которая может быть оценена по имперической ф-ции распределения.

$$\sup_{x,n\to\inf} \left| F(x) - \hat{F}_n(x) \right| \to 0$$

По теореме Гливенко-Кантелли. То есть, при увеличении объема выборки N, теоретическая и империческая ф-ции сходства совпадают. По имперической ф-ции распределения строят *выборочное среднее* или *империческое среднее* (выборочный аналог 1-ого начального момента, или мат. ожидания)

$$x_{-} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Выборочная дисперсия:

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - x^-)^2$$

Выборочные моменты поряда r:

$$\mu_{r,a} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r\right)^{1/r}$$

В ряде случаев, требуется оценить *размах выборки* (разность между наибольшим и наименьшим значениями результатами наблюдений):

$$R_n = \left|x^{(n)} - x^{(1)}\right|$$

Иногда дисперсия не является особо информативной, и потребуется размах. Либо, возможна обратная ситуация.

3 Точечные оценки параметров

Пусть имеется некоторая сл. вел. $\xi : F(x, \Theta), f(x, \Theta), \Theta$ – параметр ф-ции распределения. f – плотность распределения с.в, F – ф-ция распределения с.в.

Пример: $f(x, \lambda) = 1 - e^{x\lambda}$

В случае

$$F = \begin{cases} F(x, \Theta_1) \\ F(x, \Theta_2) \\ F(x, \Theta_3) \end{cases}$$

Обычно говорят о параметрическом семействе распределений, в котором Θ принимает различные значения.

Вводят ф-цию от рез-тов наблюдений: $\phi = \phi(x_1, x_2, .., x_n)$, где аргументы – это результаты наблюдений, называемую *статистикой*. Задача построения точечной оценки параметра Θ сводится к нахождению значения статистики, такой что:

$$\hat{\Theta} = \Theta(x_1, x_2, \ldots, x_n)$$

$$\sup_{n\to\infty} |\hat{\Theta} - \Theta| \to 0$$

Необходимо установить *эффективную оценку*, рекомендуемую в качестве результата.

3.1 Свойство оценок

$$\hat{\Theta} = \Theta(x_1, x_2, \ldots, x_n)$$

$$\lambda = \frac{1}{\overline{x}} = \frac{1}{\frac{1}{n}\sum\limits_{i=1}^{n} x_i} = \frac{n}{x_1 + \ldots + x_n}$$

Оценка $\hat{\Theta}$ является не смещенной оценкой пар-ров Θ, если ее мат. ожидание $M(\hat{\Theta})$ совпадает с параметром Θ, т.е. $M(\hat{\Theta}) = \Theta$. Оценка называется *асимптотически не смещенной*, если $\lim\limits_{n\to\infty} M(\hat{\Theta}_n) = \Theta$.

Таким образом, на св-во оценок влияет объем выборки. Если $\lim\limits_{n\to\infty} P\{|\hat{\Theta}_n - \Theta| < \epsilon\} \to 1$, то говорят, что величина сходится по вероятности.

Пусть $\hat{\Theta}_n$ – асимптотически не смещенная оценка параметра n, и то что ее дисперсия

стремится к нулю $\lim\limits_{n\to\infty} S^2(\hat{\Theta}_n) \to 0$, то эта оценка является *состоятельной*. Таким образом, асимптотическая несмещенность оценки Θ, и минимизация разброса значения пар-ра, при $n \to \infty$ обеспечивает состоятельность оценки (Теорема приводится без доказательства).

Если $S^2(\hat{\Theta}_n^1) = M(\hat{\Theta}_n^1 - \Theta)^2 \leq M(\hat{\Theta}_n^2 - \Theta)^2 = S^2(\hat{\Theta}_n^2)$, то оценка $\hat{\Theta}_n^1$ является эффективной по сравнению с $\hat{\Theta}_n^2$

3.2 Методы построения точечных оценок

Есть n объектов, и столько же будет пар изменяемых характеристик $\xi_1 \dots \xi_n$ Может стоять задача определения связей между не наблюдаемыми одновременно параметрами

Рис. 2 – Тренд линия

Различают 2 случая:

1. Моделирование производят на выборках ξ_1, ξ_2.

2. Не наблюдаемых одновременно параметрах.

В 1 случае рассматривают пары точек $(x_i, y_i), i = \overline{1, n}$, характеризующие $\xi_1 : \{x_1, \dots x_n\}$ $\xi_2 : \{x_1, \dots, x_n\}$ – случайные выборки. Объемы выборок одинаковые. Как правило, для определения точечных оценок пар-ров ф-ции 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 2-мя случайными переменными. Термин *стохастическая связь* был введен Чупровым в 1926 году. *Стохастическая связь* — это такая связь, при которой значению величины $x \in X$ соотвествует одно или более значений $y_1, \dots, y_k \in Y$. Таким образом, получение с.в. y_i изменяет вероятность появления других значений, но не обеспечивает их появление. Именно поэтому говорят, что стохастическая связь не является причинной.

(б) Ф-циональная связь описывается зависимостью, в которой с.в $x \in X y \in Y$, соответствует только одно значение $y \in Y$. Функциональная связь причинна, т.е. для каждого x_i соответствует конкретная реализация y_i. Предполагает взаимоодназначное преобразование.

Как правило, на практике исследователи имеют дело с вариантом В. Присутствуют ошибки измерения, имеет место разброс реализации остостиеельно некоторой ф-циональной зависимости в той или иной степени достоверно описывающей входные данные. Такую ф-цию называют *ф-цией тренда*. В случае достаточно тесной стохастической связи, задача сводится к выделению тренда и его анализа. Задачу В решают с помощью метода наименьших квадратов.

В случае 2 (Не наблюдаемых одновременно выборок), В этом случае, размер выборки может отличаться

$\xi_1 : \{x_1, \ldots, x_n\}$,

$\xi_2 : \{y_1, \ldots, y_m\}$,

3.2.1 Метод моментов

Пусть $[x_1, x_2, \ldots x_n]$ — независимая случайная выборка из генеральной совокупности с ф. распр. $F(x, \Theta)$, и плотностью распределения $f(x, \Theta)$, где Θ – параметр распределения.

Можем посчитать теоретическое зн-е мат. ожидания:

$$\mu(x) = \int\limits_{-\infty}^{\infty} x f(x, \Theta) dx = \frac{1}{n} \sum_{i=1}^{n} x_i \to \hat{\Theta} \tag{1}$$

Формула 1 позволяет получить оценку неизвестного параметра распределения. В случае многомерного параметра:

$$\vec{\Theta} = (\Theta_1, \ldots \Theta_n)$$

Используют r первых моментов:

$$\mu_r = \int\limits_{-\infty}^{\infty} x^r f(x, \overline{\Theta}) dx$$

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

$a = 0$ – момент начальный

$a = \overline{x}$

В общем случае $b = 1$. Если рост величины момента значителен, исследователь использует нормализацию. Можно выбрать $b = 1/r$

$$\begin{cases} \mu_1 = \int\limits_{-\infty}^{\infty} x f(x, \Theta) dx = \overline{x_1} = \hat{\mu_1} \\ \mu_2 = \int\limits_{-\infty}^{\infty} x^2 f(x, \Theta) dx = \overline{x_2} = \hat{\mu_2} \\ \qquad\qquad \ldots \\ \mu_r = \int\limits_{-\infty}^{\infty} x^r f(x, \Theta) dx = \overline{x_r} = \hat{\mu_r} \end{cases}$$

В случае 2х параметров, рассматривают 1-ый начальный момент (мат. ожидание и

выборочное среднее), и 2ой центральный момент (дисперсия и выборочная дисперсия).

В случае, оценки ф-ции связи между 2мя наблюдаемыми пар-рами, используют модифицированный метод моментов.

В предположении, что между с.в. ξ_1, ξ_2 имеется достаточно тесная стохастическая случайная зависимость, и она м. быть описана функциональной зависимостью (линейная зависимость) вида:

$$\xi_1 = \phi(\xi_2) = k\xi_2$$

В этом случае, оценка k будет получена из:

$$\hat{\mu}_{\xi_1} = \phi(\hat{\mu_{\xi_2}})$$

Пример: $\xi_1 = (100, 200, 300)$ $\xi_2 = (10^5, 3 \cdot 10^5, 5 \cdot 10^5)$

$\hat{\mu_{\xi_1}} = \overline{x_1} = 200$

$\hat{\mu_{\xi_2}} = \overline{x_2} = 3 \cdot 10^5$

$\hat{\mu_{\xi_1}} = \phi(\hat{\mu_{\xi_2}}) = k\hat{\mu_{\xi_2}} \to 2 \cdot 10^2 = k \cdot 3 \cdot 10^5 \Rightarrow \hat{k} = 6,7 \cdot 10^{-4}$

(3*)

$y = \alpha x^\beta = \phi(x)$

$\xi_1 = \alpha \xi_2^\beta$

Между ξ_1 и ξ_2 существует стохастическая зависимость в пр. сл.

$$\hat{\mu}_{\xi_1} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\alpha}{m}\sum_{j=1}^{m} x_i^\beta = \hat{\mu_{\xi_2}}$$

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

$$\ln(\alpha) - \beta \ln(\frac{1}{n}\sum_{j=1}^{n} x_i) \to \beta A + \ln \alpha = A \qquad (4*)$$

Так как нейзвестны парамеры α и β, необходимо построить второе уравнение.

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

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu_{\xi_1}}) = \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i^\beta - \alpha\hat{\mu_{\xi_1}})^2 = \hat{S_{\xi_2}^2} \qquad (5*)$$

Из 5* получают 2-ое уравнение, замыкая систему

$$(6*) = \hat{\alpha_i}\hat{\beta} = \begin{cases} \hat{\mu_{\xi_1}} = \hat{\mu_{\xi_2}} \\ \hat{S^2_{\xi_1}} = \hat{S^2_{\xi_2}} \end{cases}$$

(6*) СЛАУ, во избежание нелинейности, (5*) также лианеризуют Т.о модификация метода моментов оказывается индивидуальным для любого вида ф-ции связи. Кроме того, необходимо обосновать выбор ф-ции связи.

Обоснование вида ф-циональной зависимости по рез-татам эксперимента, производят на основе анализа выборочных ф-ций распределения.

Имеем:

$F(x, \alpha_1, \beta_1) = 1 - e^{-\alpha_1 x^{\beta_1}}$ $F(y, \alpha_2, \beta_2) = 1 - e^{-\alpha_2 y^{\beta_2}}$

Выполняем замену:

$1 - e^{-\alpha_1 x^{\beta_1}} = 1 - e^{-\alpha_2 \phi(x)^{\beta_2}}$

$\alpha_1 x^{\beta_1} = \alpha_2 \phi(x)^{\beta_2} \Rightarrow \frac{\alpha_1}{\alpha_2} x^{\beta_1} = \phi(x)^{\beta}_2 \Rightarrow \phi(x) = \frac{\alpha_1}{\alpha_2} x^{\beta_1/\beta_2} \rightarrow$

Где:

$\hat{\alpha} = \frac{\alpha_1}{\alpha_2}$

$\hat{\beta} = \frac{\beta_1}{\beta_2}$

Т.о. подход к точечной оценки пар-ров ф-ции связи, имеет следующие шаги:

1. Анализируются выборки, соответствующие исследуемым велечинам, с целью выявления выбросов;

2. Строятся выборочные ф-ции распределения вероятности. Выборочные ф-ции описываются теоретическим распределением;

3. На основе вида ф-ции распределения, анализируются зависимость между исследуемыми переменными. Результатом шага является обоснование выбора ф-ции связи. Параметры при этом остаются неизвестными;

4. Анализируются количество пар-ров. Рекомендуется приближать много- параметрические ф-ции, 1-2х параметрические;

5. Число уравненеий в системе совпадает с числом параметров, что позволяет выбрать такое же кол-во выборочных моментов для обеих выборок. И реализовать

модифицированный метод моментов для получения пар-ра ф-ции связи между исследуемыми переменными.

Преимуществом стохастического моделирования перед аналитическим – высокая степень формализация к оценке ф-циональной зависимости между хар-ками системы, а также алгоритмообоснование степени влияния того или иного явления на рез-т исследований.

К недостаткам метода можно отнести феноменологический пар-р модели (те, которые построены по рез-татам эксперимента, и зависят от них). Второй недостаток – существенное упрощение вида ф-ций связи на каждом этапе моделирования.

Несмотря на субъективность выбора ф-ций связи, стохастические модели (параметры и состояния которых — представлены случайными величинами) имеют незначительную вычислительную погрешность, что в некотором смысле компенсирует недостатки, связанные с упрощением.

3.2.2 Метод максимального правдоподобия

Пусть $x_1, x_2, \dots x_n$ из X – генеральной совокупности с ф. распределения $F(x, \Theta)$, и $f(x, \Theta)$. В МС, в основе методов лежит выбор ф-ции, зависящей от всех эл-ов выборки, и называемой – статистикой.

В ММП, в качестве статистики выбирается ф-ция *совместная плотность распределения*:

$$f(x_1, \Theta) \cdot f(x_2, \Theta) \cdot \dots \cdot f(x_n, \Theta) = L(x_1, x_2, \dots, \Theta)$$

L – называется *ф-ция максимального правдоподобия*. Оценка максимального правдоподобия находится из соотношения (далее «$\to \hat{\Theta}$» имеется в виду что равенство нулю будет выполнено при «$\Theta \to \hat{\Theta}$»):

$$L(x_1, \dots, x_n, \hat{\Theta}) = \max L(x_1, \dots, x_n, \Theta)$$
$$\frac{dL(x_1, ..., x_n, \Theta)}{d\Theta} = 0 \to \hat{\Theta}$$

Если L принимает существенные зн-я, чаще:

$$\frac{d\left(\ln L(x_1, ..., x_n, \Theta)\right)}{d\Theta} = 0 \to \hat{\Theta}$$

В случае многомерного параметра $\Theta = (\Theta_1, \ldots, \Theta_p)$ необходимо решить СЛАУ в общем случае, система может быть нелинейной. При фиксированном $x_1, ..., x_n$:

$$(7*) = \hat{\overline{\Theta}} = (\hat{\Theta}_1, \ldots, \hat{\Theta}_p) \begin{cases} \dfrac{dL(x_1, ..., x_n, \overline{\Theta})}{d\Theta_1} = 0 \to \hat{\Theta} \\ \dfrac{dL(x_1, ..., x_n, \overline{\Theta})}{d\Theta_2} = 0 \to \hat{\Theta} \\ \dfrac{dL(x_1, ..., x_n, \overline{\Theta})}{d\Theta_p} = 0 \to \hat{\Theta} \end{cases}$$

Достоинством такого подхода является получение асимптотически несмещеннных, и асимптотически эффективных оценок, при $n \to \infty$. При этом, невозможен нелинейный вид (7*), что приводит к накоплению выч. погрешности.

4 Доверительные интервалы

$$x = \{x_1, x_2, ..., x_n\} \in X_n$$

$$F(x, \Theta), f(x, \Theta)$$

Предположим, что для оцениваемого пар-ра Θ, построим интервал $\Theta \in [\Theta_-, \Theta^-]$. Интервал $[\Theta_-, \Theta^-]$, называется *доверительным интервалом* с вероятностью попадания γ. Здесь γ есть величина $P\{\Theta_- \leq \Theta \leq \Theta^-\} = \gamma$

4.1 Методы построения доверительных интервалов

Статистикой метода называется любая выборка $T = T(x_1, x_2, \ldots, x_n)$

Если ф-ция зависит от параметра Θ, т.е. $T = T(x_1, x_2, \ldots, x_n, \Theta)$ то говорят о *центральной статистике.* При выборка $x_1, x_2, ...x_n$ – независимая случайная повторная, параметр Θ – скалярная величина, но в общем случае может рассматриваться как вектор. Ф-ция T является монотонной относительно пар-ра Θ

Квантилем уровня α ф-ции распределения $F(x, \Theta)$. Квантиль обозначим как K. Тогда, $F(K) = \alpha$. Можно дополнительно дописать $F(K) = P\{x \leq K\}$.

РИС (Про квантиль уровня p)

Зададимся 2-мя малыми числами ϵ_1, ϵ_2, и опередим k_1, k_2 это будут квантили уровней

$$P\{x \leq k_1\} = F(k_1) = \epsilon_1$$

$$P\{x \leq k_2\} = 1 - F(k_2) = \epsilon_2$$

Откуда, получаем:

$$P\{x > k_1\} = 1 - \epsilon_1$$

$$P\{x \leq k_2\} = 1 - \epsilon_2$$

Тогда, получаем что γ будет равна

$(*)P\{k_1 < T(x_1, x_2, ..., x_n, \Theta\} \leq k_2\} = \gamma$

В (*), левая граница интервала должна быть замкнутой, т.е.

$P\{k_1 \leq T(x_1, x_2, ..., x_n, \Theta\} \leq k_2\} = \gamma$

Это допущение возможно в силу монотоности возмущения $F(x, \Theta)$

$F(k_2) - F(k_1) = 1 - \epsilon_2 - \epsilon_1$

Т.о., $\gamma = 1 - \epsilon_1 - \epsilon_2$

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

Далее, нижнаяя и вехрхняя границы Θ (Θ_-, Θ^-) определяются как мин. и макс. значение среди всех возможных значений пар-ра Θ, удовлетворяющих (**).

Если центральная статистика монотонно возрастает, то границы доверительного интервала находят из системы

$$(3*) \begin{cases} T\{x_1, x_2, ..., x_n, \Theta\} = k_1 \\ T\{x_1, x_2, ..., x_n, \Theta\} = k_2 \end{cases}$$

$$(4*) \begin{cases} T\{x_1, x_2, ..., x_n, \Theta\} = k_2 \\ T\{x_1, x_2, ..., x_n, \Theta\} = k_1 \end{cases}$$

Чаще статистику выбирают таким образом.

Пример доверительного оценивание интревала (для случая нормального распределения).

$N(\mu, \sigma)$ – исследуемая ф-ция распределения вероятностей

$N_{0,1} = N(0, 1)$ – ф-ция распределения значений статистики.

При известной дисперсии для доверительного оценивания мат. ожидания, в качестве исходной центральной статистики, используется следующая

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

Задача многопараметрического оценивания на несколько порядков (сложность повышается на порядок при каждом новом неизвестном параметре, вычислительно сложна). В силу чего, исследователи сводят задачу многопараметрического оценивания к задаче 1о параметрического оценивания.

Вид статистики (5*) выбирается исследователем, обладает следующими св-вами:

1. Монотонно убывающая по μ ф-ция;

2. имеющая стандартное нормальное распределение.

Вводится $\epsilon = 1 - \gamma/2$, и квантиль уровень ϵ: K_ϵ

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

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

5 Лабораторная работа 1 (Разбор)

Рассмотрим понятие убытка

$$U = P - R$$

U – убыток,

P – планируемый доход,

R – реальный доход метрополитена.

Как определить P:

1. Задать, например 100 млн. крон/месяц.
2. $P = ax_1^*$,

   a – средняя стоимость поездки. x_1^* – все воспользовались за месяц метрополитеном.
3. $P = \sum_{i=1}^{k} a_i x_1^{(i)^k}$,
   $\sum_{i=1}^{k} x_i^{(i)^k} = x_1^*$
   $P = f_1(x_1^*) = \phi(x_1)$

Существует аналитическая зависимость между планируемым доходом, и количеством входящих в метрополитен в течение месяца

В этом случае, можно предположить что между планируемым доходом и кол-вом пассажиров, вошедших на конечных станциях, существует тесная стохастическая связь между планируемым доходом и числом пассажиров, вошедших на конечной станции

Необходимо подтвердить или опровергнуть наличие тесной стохастической связи в предельном случае, описываемой $P = \varphi(x_1)$

1. $x_1^* = kx_1$ – грубая оценка;
2. $x_1^* = \xi(x_1) \to \varphi(x_1)$;

   ...

Возможна грубая оценка (на основе Центральной предельной теоремы) общего числа входящих в метрополитен в течение месяца (года).

В качестве примера, $x_1^* = kx_1$, коэффициент k – количество станций, а x_1 – число вошедших. Альтернативой является оценка стохастической связи $x_1^* = \xi(x_1) \to \varphi(x_1)$, что требует проведения натурного эксперимента (или полунатурного), 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. $P = g(x_2^*) = g(\xi_2(x_2)) = \varphi(x_2, \xi_3(x_3^*))$

   В предположении, что между доходом, и количеством проданных билетов имеется ф-циональная зависимость $P = ax_2^*$, а между билетами, проданных в метрополитене и билетами, проданными автоматами на конечных станциях метро, существует стохастическая связь вида:

$$x_2^* = \xi_2(x_2)$$

Так как, билет могут быть проданы вне метрополитена, а также использоваться в любой день после продажи, в т.ч. они могут сгореть (стать просроченными), связь между количеством проданных билетов во всем метрополитене, и кол-вом прокомпостированных билетов в конкретный день исследования, является стохастической.

Необходимо подтвердить, или опровергнуть наличие тесной связи между переменными x_2^*, x_3^*, и оценить зависимость:

$$P = \varphi_2(x_2, x_3)$$

Теперь, можем записать, что:

$$U = \varphi_1(x_1) - \varphi_(x_2, x_3)$$

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

Уже имея выбранные связи, строим зависимости

$$U(x_1, x_2, x_3) = \varphi_1(x_1) - \varphi_2(x_2, x_3)$$

К следующему разу, необходимо:

1. Реализовать процедуру установления ф-ции связи между двумя ненаблюдаемыми одновременно пар-рами

$$F(x_1, x_2, x_3) = \hat{F}(x_1) - \hat{F}(x_2, x_3)$$
$$f_1 = \hat{F}(x_1)$$
$$f_2 = \hat{F}(x_2)$$

2. Протестировать на любых наборах данных

Имеется выборка для которого посчитано среднее значение $_{i_1}$.

$$x_1, x_2, \ldots, x_{n_1}$$

Также имеется независимая выборка со средним значением y_{i_2}.

$$y_1, y_2, \ldots, y_{n_2}$$

Имеются также *контрольные* ξ_1^K, ξ_2^K и *рабочие выборки* ξ_1^P, ξ_2^P.
По рабочим выборкам, необходимо построить модель $\xi_2 = \varphi(\xi_1)$

$$\xi_1 = \alpha \xi_2^\beta$$

$$\begin{cases} \xi_1^P = \alpha \xi_2^{P\beta} \Rightarrow \hat{\alpha}, \hat{\beta} \\ \xi_1^K \ ? \ \hat{\alpha} \xi_2^{K\hat{\beta}} \end{cases}$$

6 Основные понятия проверки статистических гипотез

Статистическая гипотезой H называют любое утверждение, относительно ф-ции распределения $F(x)$ с.в. x, касающееся ее вида, и значения ее пар-ров.

Гипотезы H проверяются путем сопоставления выдвинутых предположений с результатами экспериментов, которые в статистике представляю собой N независимых, повторных случайных наблюдений над с.в. x.

РИС 5.1

Ф-ция $G(x)$ – приближенный аналог. Затем можно получить семейство аналогов:

$$G_i(x),\ \ x = \overline{1,n}$$

Различают две постановки задачи:

1. Теоретическая ф-ция распределения **считается** известной. В этом случае, проверяемая гипотеза называется *простой*
2. В другом случае, теоретическая ф-ция распределения неизвестна, и проверку стат. гипотезы H осуществляют для семейства $G_i(x)$ – ф-ций распределений, отличающихся значениями пар-ров.

В обоих случаях, гипотезу проверят на основе статистических критериев.

Мы будем обозначать:

- H_O – основную гипотезу
- H_A – альтернативную

Гипотеза, справедливость которой проверяется — называется *основной*. В зависимости от того, какие альтернативы основной гипотезе возможны в предметной области, формулируют альтернативную гипотезу.

Статистический критерий — совокупность правил, позволяющих по полученной выборке принять основную гипотезу, и отвергнуть альтернативную; или наоборот.

Имеется общий принцип построения статистических критериев.

1. Задается некоторая ф-ция $S = S(x_1, x_2, \ldots, x_n)$ от всех элементов выборки, называемая *статистикой критерия*;

2. Множество Ω Разбивается на 3 подмножества $T_o, T_{kp}, T_u \in \Omega$:

   (a) T_o – множество принятия решений (соответствует основной гипотезе);

   (b) T_{kp} – критическое множество (соответствует альтернативной гипотезе). Если критерий таков, что критическим является как малые, так и большие значения статистики, то критерий называется *двусторонним*. В противном случае – односторонний;

   (c) T_u – Множество индифферентности, отделяющее основную от критической.

3. Если $S \in T_o$, то H_0

   Если $S \in T_{kp}$, то H_1.

В силу того, что ф-ция $S = S(x_1, x_2, \ldots, x_n)$ является случайным, событие $S \in T_{kp}$ может произойти как при справедливости H_0, так и при справедливости H_A.

Ошибкой I рода — называется принятие альтернативной гипотезы H_A, когда верна H_0. α – вероятность этой ошибки.

$$\alpha = P(S(x_1, x_2, \ldots, x_n \in T_{kp})|H_0)$$

Ошибкой II рода — назывется принятие основной гипотезы, когда верна альтернативная. β – вероятность этой ошибки.

$$\beta = P(S(x_1, x_2, \ldots, x_n \in T_O)|H_A)$$

$\alpha,\ \beta$ — незначительные величины. Цель илсследователя – минимизировать эти значения, но одновременно эти величины не минимизируются.

Вводят $\gamma = 1 - \beta$ – мощность критерия.

α задается исследователем, при этом величина γ максимизируется.

α называют *размером критерия*. Как правило, малое значения α не задают маленькой, так как размер критерия будет малым. β – уровень *значимости критерия*.

Если известен закон распределения с.в., статистические выводы оказываются достаточно точными. В тоже время, необходимо проверить, соответствуют ли экспериментальные данные этому распределению. Для проверки используют *критерий согласия*, в частности *критерий Колмагорова-Смирнова*, и *критерий* ω^2.

6.1 Критерей Колмагорова-Смирнова

Критерий основан на статистике Колмагорова. Теоретическая и выборочная ф.р. сравниваются в некоторой равномерной метрике, которая формулируется следующим образом:

$$D = \sup_x |F(x) - F_n(x)| \tag{2}$$

$$H : G(x) = F(x)$$

$$G(x) \ ? \ F_n(x)$$

Формула 2 называется *статистикой Колмагорова.*

РИС 5.2

Мы будем искать максимальное отклонение между теоретической ф-цией распределения и ее аналогом.

Критерий Колмагорова-Смирнова (КС)

Если гипотеза H верна, и при $n \to \infty$ выполнено: $F_n(x) \to G(x)$, и $G(x) \neq F(x)$, то $F_n(x) \to F(x)$ при $D \leq D_\beta$, где D выбирается из процентных точек, как величина на пересечении строки n и столбца β.

При $n > 35$, D_β вычисляется:

$$D_\beta = \sqrt{-0.5 \ln \beta}$$

$$n > 35$$

Статистический критерий проверки гипотезы называют *состоятельным против альтернативы* H_A, если вер-ть отвергнуть H_O, когда на самом деле верна H_A стремится к единице, при неограниченном росте n.

Состоятельность критерия КС означает, что любое отличие распределения выборки от теоретической, будет обнаружено, если измерения продолжаются бесконечно долго, что обеспечивается не всегда. Альтернативой является применение *критерия* ω^2.

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

На практике пользуются *эмпирической оценкой* w_n^2 (см. 3)

$$nw_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(F(x_i) - \frac{2i-1}{2n} \right)^2 \tag{3}$$

Величина nw_n^2 не должна превышать табулированного порогового значения.

7 Методы многомерного статистического анализа (МСА).

Каждый объект выборки может содержать наблюдения более чем над 1-ой случайной переменной (задача множественной регресии). Различают две постановки задачи:

1. $\xi_1 = \varphi(\xi_2, \xi_3, \ldots, \xi_n)$, где ξ_1 – зависимая переменная, $\xi_2, \ldots, \xi_n$ – независимые переменные.

2. $\{\xi_2, \xi_3, \ldots, \xi_n\}$ — рассматривается вектор независимых случайных величин, имеющих многомерное распределение.

Многомерный анализ — анализ множественных измерений свойств случайной выборки, для анализа которых используются группа статистических методов.

В рамках многомерного статистического анализа (МСА), ставятся следующие задачи:

1. Задача стохастического моделирования;

2. Исследования структуры сложной системы или ее модели, в том числе описания поведения системы во времени, (изменение тренда, описывающего поведение системы, поиск периодических (или квази-периодических) колебаний, оценка задержек (лагов), и т.д.);

3. Прогнозирование. Задачи прогнозирования будущего развития процесса;

4. Анализ взаимодействия между процессами;

5. Прогнозирование 2-х и более процессов.

Для решения этих и аналогичных задач, используются анализ временных рядов, который включает методы:

- корреляционного анализа;
- спектрального анализа;
- методы авторегрессии и скользящего среднего.

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 относятся вновь поступивший объект.

Для решения задач классификации с 2-мя группами, как правило используется дискриминантный анализ.

Дискриминантный анализ — предполагает построение дискриминирующей ф-ции, аргументами которых являются измеряемые величины. Далее, выделяются области, при попадании в которые объект относится к первой, или второй группе. Если требуется разделение на 3, и более групп, и/или нет сведений о характеристиках объектов, определяющих группу, используются методы кластерного анализа.

Метод кластерного анализа — предполагает идентификацию ядер кластера (измеряемые характеристики являются координатами точек в гиперпространстве). Оценивается близость вновь прибывшей точки до ядра каждого кластера

Возможна ситуация, когда характеристики объектов не измеряемы (возможны качественные оценки), или количественные оценки не информативны, тогда используют задачи шкалирования (ранжирования). Такая группа методов называется *методы шкалирования.*

8 Метрики $\rho(x', x'')$.

Понятие расстояния между объектами отражает меру сходства объектов между собой по всей совокупности используемых признаков.

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

1. $d_{ij} \geq 0$ — неотрицательность расстояния;
2. $d_{ij} = d_{ji}$ – симметричность;
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 расстояния, чаще применяют термин *метрика*. В математике, наиболее распространенной и универсальной, является *метрика Миньковского* (N признаков у объекта).

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n} \left(x_i^k - x_j^k\right)^\lambda}$$

Разница между количественными характеристиками признаков i и j-ого объекта определяется следующим образом:

$$\left(x_i^k - x_j^k\right)$$

Для $\lambda = 1$, $d_{ij} = \sum\limits_{i=1}^{n} \left|x_i^k - x_j^k\right|$. Модуль в метрике вводится для независимости от знака. Эта метрика также носит название *расстояния городских кварталов*, или *расстояние Манхетена*.

Для $\lambda = 2$, $d_{ij} = \sqrt{\sum\limits_{i=1}^{n} \left(x_i^k - x_j^k\right)^2}$ – метрика Евклида.

Для $\lambda = \infty$, $d_{ij} = \max\limits_{k} |x_i^k - x_j^k|$ – метрика Чебышева.

Существуют иные метрики, принципиально отличающиеся от метрики Миньковского.

Например, метрика Канберры (см. формулы 4, 5):

$$d_{ij} = \frac{\sum_{k=1}^{n} |x_i^k - x_j^k|}{\sum_{k=1}^{n} |x_i^k| + \sum_{k=1}^{n} |x_j^k|} \tag{4}$$

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_i^k - x_j^k|}{|x_i^k| + |x_j^k|} \tag{5}$$

Метрика Канбера обычно используется для данных, измеряемых в диапазоне [0,1].

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

8.1 Оценка качества кластеризации.

Оценка качества кластеризации может быть проведена следующим образом:

1. Вручную на тестовых данных;
2. Установление контрольных точек, и проверка на полученных кластерах;
    - Контрольные точки интуитивно выбираются исследователем (признаки отчисляемого и не отчисляемого после 1-ой сессии студента). Набор контрольных точек ограничен, а выбор субъективен.
3. Добавление в модель новых переменных.
    - Если при введении дополнительного признака, кластеризация стабильна, то это свидетельствует о высоком качестве кластеризации.
4. Создание и сравнение кластеров с исользованием различных методов.
    - Различные методы продуцируют разные наборы кластеров, но если в целом результаты сходны, можно говорить о высоком качестве кластеризации.

8.2 Методы кластерного анализа.

8.2.1 Метод K-средних (K-means)

Как правило, его описывают как иерархический. Кластеры представлены в виде центроидов, являющихся «центром масс» всех объектов, входящих в кластер.

В отличие от классических иерархических подходов, которые не требуют предварительных предположений относительно числа кластеров, в алгоритме K-средних необходимо проверить гипотезу о наиболее вероятном количестве кластеров (3 или 2).

Алгоритм строит K кластеров на возможно больших расстояниях друг от друга. Т.е. кластеры, должны быть как можно более удалены друг от друга.

Общая идея: Заданное фиксированное число кластеров сопоставляется кластерам таким образом, что среднее значение для всех переменных, близки но значительно удалены от средних других кластеров.

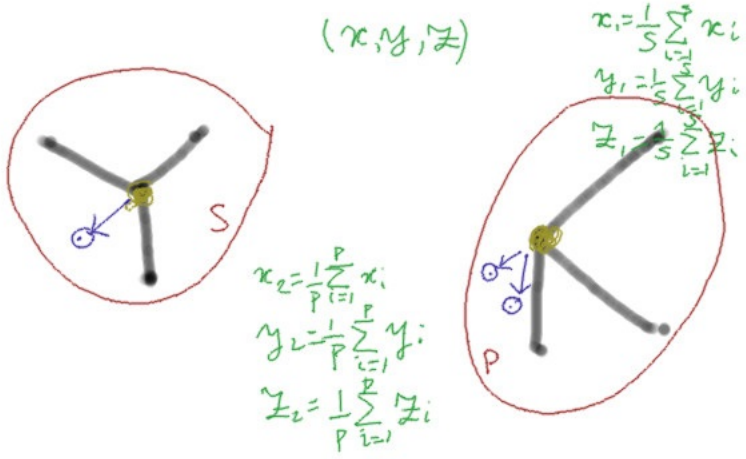


Рис. 3 – Метод K-средних, пример работы. Рассматриваются два класса s и p. Зеленым цветом отмечено вычисление центроидов классов.

Из-за смещения центра масс, на каждом новом шаге метода точки могут перемещаться из кластера в кластер, на i-ом шаге процесс стабилизируется, что свидетельствует о завершении кластеризации. Возможна ситуация зацикливания алгоритма. В этом случае, необходимо определить условие окончания счета. Если переход повторился, то кластеризацию можно считать завершенной.

$$D(X,Y) = dev(XY) - (dev(X) + dev(Y)) \tag{6}$$

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

$dev(X)$ - дивергенция.

Если в каждом k-ом кластере S элементов, то стоимость кластера:

$$c(S_k)$$

Идея такова, что мы минимизируем сумму по всем кластерам, ставится задача минимизации суммы стоимости кластеров по всем кластерам:

$$\min \sum_{i=1}^{k} c(S_i) \tag{7}$$

Функция стоимости субъективно выбирается исследователем из соображения возможности ее минимизации. Как правило, минимизируют суммы квадратов расстояний до центров кластеров (для нормализованных, стандартизированных данных), либо сумма дисперсий.

Преимущества алгоритма k-средних:

- Простота использования;
- Быстродействие;
- Простой алгоритм выбора начального приближения центров кластеров (первые k точек).

Недостатки алгоритма k-средних:

- Алгоритм очень чувствителен к выбросам;
- Плохо работает на больших объемах данных (связано это с "псевдопереобучением");
- Выбор в качестве начальных первые k-центроидов влияют негативно (на точность кластеризации, время исполнения, алгоритм не очень уверенный в реализации).

8.2.2 Метод Варда

Является *иерархически-агломеративным*. В качестве начальных центроидов выбираются все объекты выборки. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний до центров кластеров.

На каждом шаге алгоритма объединяются такие 2 кластера, которые приводят к минимальному увеличению целевой функции (функция стоимости), формула 7.

$$\begin{cases} x^{(0)} = (x' + x'')/2 \\ y^{(0)} = (y' + y'')/2 \\ z^{(0)} = (z' + z'')/2 \end{cases}$$

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

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

$$C(A) = (A', A^0) + \rho(A'', A^0)$$

Прошли по всем данных, осуществили объединение. Далее, на следующем шаге появятся новые точки, мы выполняем тоже самое только для точек, полученых на предыдущем шаге.

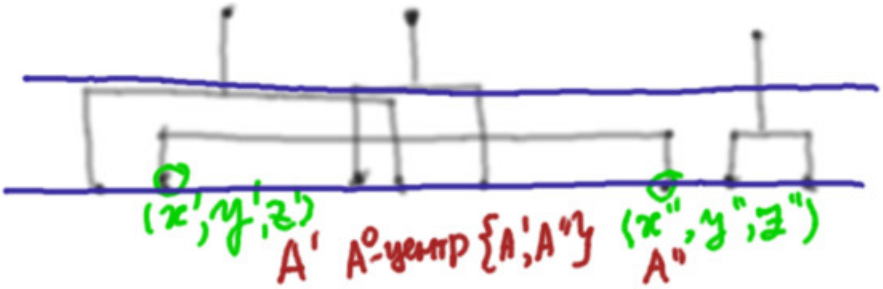


Рис. 4 – Дендрограмма

Фактически, для оценки расстояния между кластерами используются методы дисперсионного анализа. Минимальное увеличение ф-ции стоимости соответствует требованию к минимизации для групповой внутри суммы квадратов. Т.е. метод Варда обеспечивает объединение близких, сходных объектов при удаленности самих групп.

Условием окончания счета является достижения расстояния между кластерами больше заданного. Счет останавливается, при достижении нужного числа кластеров.

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

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

8.2.3 Метод ближайшего соседа.

Метод классифицируется как *иерархический, агломеративный.*[1] Расстояние между кластерами оцениваются 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 (обычно ограничивается этими задачами).

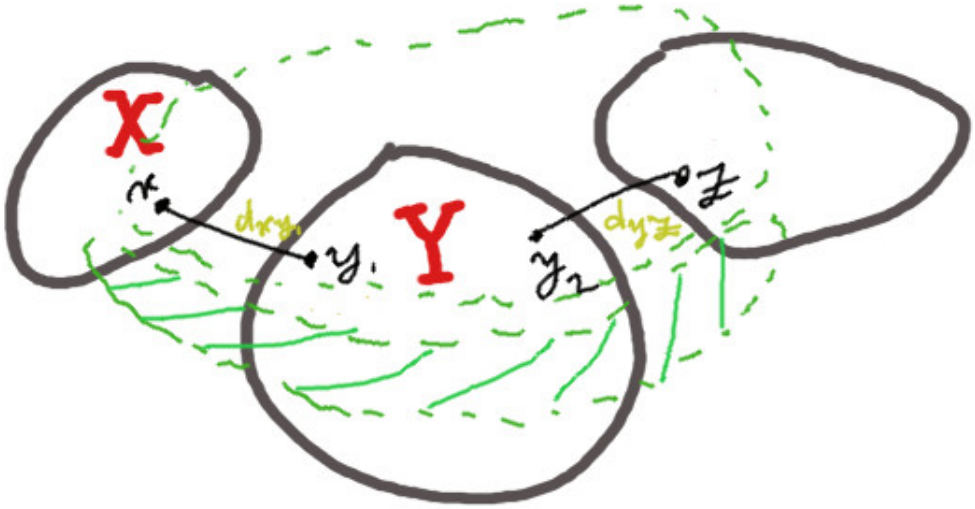


Рис. 5 – Кластеризация методом наиболее удалённого соседа.

Достоинство: позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части кластеров соеденены цепочкой (такие кластеры называют цепочечные или волокнистые).

Недостаток: структура кластера по сути определяется случайными объектами.

8.2.4 Метод наиболее удалённого соседа.

В прямом переводе иногда называют методом полной связи. Расстояние между кластерами определяется наибольшим растоянием между любыми двумя объектами в различных

[1] Она: ответь мне, только честно, да или нет, хорошо?
Он: спрашивай
Она: почему мужчины смеются над блондинками?
Он: да

кластерах:

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

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

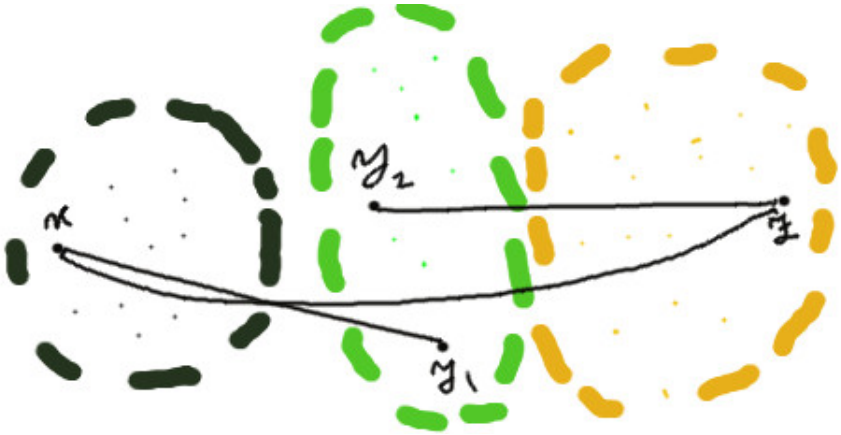


Рис. 6 – Кластеризация методом наиболее удалённого соседа.

8.2.5 Выводы.

Таким образом можно рекомендовать использовать:

- метод ближайшего соседа при описании иерархических структур;
- метод удалённого соседа для описания данных, гарантированно определяющих удалённые кластеры при тесной связи объектов внутри кластера;
- метод Варда используют с той же целью, что и метод удалённого соседа;
- когда нет предварительной информации о данных надо использовать неиерархический подход, в частности метод k-средних.

8.2.6 Лаборторная

Стобцы - автоматы. Строчки - виды билетов. В ячейках - число билетов.

Расчёт средней стоимости поездки.

$x_3 = \phi_1(x_3 1) \quad ?x_3 = \phi_2(x_3 2) \quad ?x_3 = \phi_3(x_3 3) \quad ?$

Результаты выводим в таблицу. Размерность таблицы 6x6

На основе полученных данных оценить среднюю стоимость поездки. Оценив, применимость обобщённой функции.

Лабораторная работа по Кластеризации

Необходимо объединить пары столбцов для самой первой выборки метрополитена. Таким образом будет 12*6 = 72 точки на плоскости. Нужно убедиться в том, что наибольшая точность классификации методом Вагнера будет достигаться для первого кластера (объединение столбцов 1 и 2). (Кластеризация на основе векторов 12 столбцов я сделал, но, возможно, эта задача не имеет значения)

9 Многомерный дисперсионный анализ

Дискриминантный анализ:

$$\begin{cases} x > d \\ y > ax + b, \quad a, b, d - defined \end{cases}$$

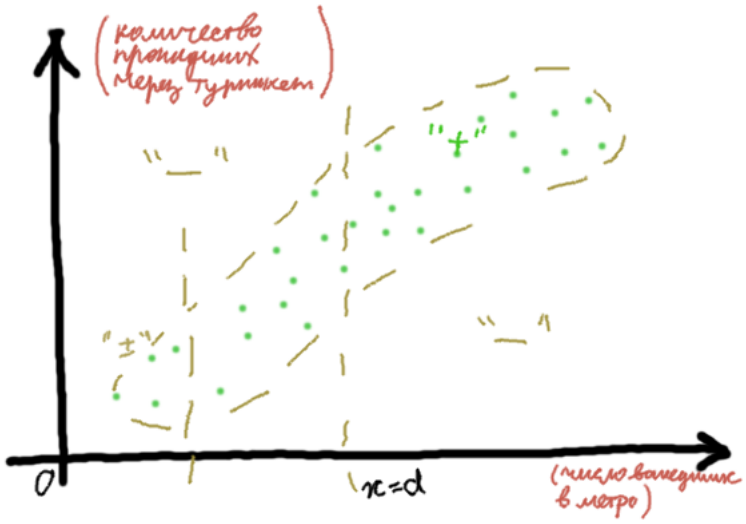


Рис. 7 – Пример работы кластеризатора для задачи «Метрополитена».

Пусть для каждого из n объектов измеряются k переменных, тогда результаты измерений могут быть представлены таблично.

x_{11}	x_{12}	$\dots$	x_{1n}
x_{21}			
$\dots$			
x_{k1}			

Таблица 1: Представление данных

Вектор-столбец x_i соответствует i-му объекту. Для каждого такого объекта можно построить линейную регрессионную модель. При объединении столбцов в матрицу можно составить многомерную линейную обобщённую модель.

$$\overline{X} = A_{1,k-1}\overline{X'_{k-1,n}} + \bar{\epsilon}_{1,n}$$

$\overline{x}$ - вектор регрессии.

$\bar{\epsilon}_{1,n}$ - вектор погрешности. Включает свободные члены и погрешности измерений.

Таким образом, задача сводится к оценке вектора неизвестных коэффициентов, и вектора погрешностей $\overline{\epsilon}_{1,n}$

В общем случае, решается переопределенная СЛАУ, т.е. количество уравнений должно превышать количество объектов.

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

Пример: пусть исследуются 16 факторов, влияющих на успеваемость студента. Требуется построить модель многомерного дисперсионного анализа позволяющую классифицировать студентов на успевающих, задолжников и неуспевающих. Для задачи был анкетирован 461 студент: 33 - успевающие, 29 - неуспевающие, 399 - задолжники.

То есть $n = 461$, $k = 16$, вектор j-го фактора $\overline{Y}_j = \{Y_{j,1}, \ldots, Y_{j,461}\}$, вектор i-го студента $\overline{Y}_i = \{Y_{1,i} \ldots, Y_{16,i}\}$. Следовательно матрица, описывающая эксперимент, имеет размерность 16×461, каждому j-му фактору, поставим в соответствие вектор (см. 8), тогда j-я строка матрицы будет иметь вид (см. 9).

$$\beta_j = \{\mu_j; \beta_{j,1}, \ldots, \beta_{j,461}\} \tag{8}$$

$$Y_j = \{1, Y_{j,1}, Y_{j,2}, \ldots, Y_{i,461}\} \tag{9}$$

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

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

$Y_{n,k}$ - все результаты экспериментов (известны),

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

$\alpha_{3,k}$ - нужно определить.

β:

- $(1, 1, 0)$ - для успевающих;
- $(1, 0, -1)$ - для неуспевающих;
- $(1, 1, -1)$ - для остальных.

$$\begin{pmatrix} Y_{1,1} & Y_{1,2} & \cdots & Y_{1,16} \\ Y_{2,1} & Y_{2,2} & \cdots & Y_{2,16} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{461,1} & Y_{461,2} & \cdots & Y_{461,16} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \\ \vdots & \ddots & \vdots \\ 1 & 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_{16} \\ \alpha_{1,1} & \alpha_{2,1} & \cdots & \alpha_{16,1} \\ \alpha_{1,2} & \alpha_{2,2} & \cdots & \alpha_{16,2} \end{pmatrix}$$

От традиционной линейной регресии перешли к свёртке используя понятие фактора.

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

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

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

$$w_{n,n} \cdot (Y_{k,n}^T)^{-1} = g_{n,k} \cdot Y_{k,n}^T \cdot (Y_{k,n}^T)^{-1}$$

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

$$\dots$$

Новые, перевычисленные параметры будут иметь следующий вид:

$$\begin{cases} \dot{Y_{k,n}}\beta n, 3 = \alpha_{k,3}^T \cdot \beta_{3,n}^T \cdot \beta_{n,3} \\ Y_{k,3}^T = \alpha_{k,3}^T \cdot \beta_{3,3}^T \\ Y_{k,3}^T \cdot (\beta_{3,3}^T)^{-1} = \alpha_{k,3}^T \cdot \beta_{3,3}^T (\beta_{3,3}^T)^{-1} \\ w_{k,3} = \alpha_{k,3}^T \cdot E_{3,3} \end{cases}$$

По результатам оценивания для фактора ночная подготовка к экзамену (в период между 22.00 и 8.00) $\mu = 1.5$, $\alpha_1 = 0.2$, $\alpha_2 = 2.6$. Далее, начинается идентификация 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, отрицательное, и нейтральное. При этом, межвидовая дисперсия факторов определяется заранее выбором веса фактора, а внутривидовая дисперсия разброса – разницой между абсолютными значениями $x_{i,j}$ для каждого j-ого фактора.

В качестве продолжения лабораторная работы, требуется:

1. Найти для каждой пары векторов метрополитена:

$$x_3 = \varphi_1(x_F)$$

$$x_3 = \varphi_1(x_D)$$

$$x_3 = \varphi_1(x_L)$$

2. Находим вероятности продажи полного (F), льготног (D), долговременного (L) билета. Здесь N – общее число билетов

$$P_F = \frac{\sum\limits_{i=1}^{N} x_F^i}{\sum\limits_{i=1}^{N} x_3^i}$$

$$P_D = \frac{\sum\limits_{i=1}^{N} x_D^i}{\sum\limits_{i=1}^{N} x_3^i}$$

$$P_L = \frac{\sum\limits_{i=1}^{N} x_L^i}{\sum\limits_{i=1}^{N} x_3^i}$$

$$P_F + P_D + P_L = 1$$

$$P_F n_F * 24 + P_D n_D * 36 + P_L n_L * 72 = S$$

$$\hat{C} = S/N$$

3. Отношение планируемого дохода / к реальному в 1012 году (среднее значения за 1 день, полученный на основе среднего от измерений за 1 месяц):

   28 млн / 2 млн.

   28-40/14 м.

   Ожидаемый поток: 1.7 млн. $\cdot$ 16 крон = 27.2 млн. (тогда стоимость поездки составляла 16 крон).

   Оценка входа x_1 относительно x_2 – числа прошедших через турникет вычисляется по формуле: $x_1 = \psi(x_2)$

   Планируемый доход: $x_1 \cdot 16$ крон

   Реальный доход: $x_2 \cdot \hat{C}$

Убыток: разница между планируемым и реальным доходом

Отчет по метрополитену должен содержать результаты работы кластеризатора (Описать что такой метод впринципе подходит для восстановления данных, где в каждый кластер будет описывать станцию), а также пункты 1-3 из текущей лекции с полученными и проанализированными результатами.

Также, необходимо оформить отчет (с формулами вычислений α, β) по предыдущей работе, где в качестве данных использовались собственные данные (не метрополитен),

17.7 — разность планируемого и реального, если C = 12.

10 ANOVA – дисперсионный анализ

ANOVA – означает **AN**alysis **O**f **VA**rialbes

Метод предложен Фишером. Согласно этим изменениям, вектор анализируемых данных $\overline{y}$ выражается как

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{10}$$

y_1	y_2	$\dots$	y_n

Таблица 2: Представление вектора $\overline{y}$

И формула 10 представима в виде суммы, связанных с результатом действия различных классификационных факторов. Пусть на результат влияют два фактора: a и b. Тогда, одномерный вектор $\overline{y}$ представим в виде таблицы $N \cdot M$, где N – число объектов, а M – число экспериментов.

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

Элементы с индексом «$\cdot$» рассматриваются как:

- $y_{i\cdot}$ — по столбцам;
- $y_{\cdot j}$ — по строкам;
- $y_{..}$ — по всем элементам таблицы.

Если величина $\{AB\}$ оказывается незначительной, то говорят об отсутствии взаимодействия факторов A и B. Аналогичное, близкое к нулю значение $\{A\}$ или $\{B\}$ говорит о слабом влиянии фактора A на результат.

Пример:

$m = 20; \quad \overline{x} = 29.23; \quad s_1^2 = 5.26$

$n = 10; \quad \overline{y} = 27.56; \quad s_2^2 = 2.19$

В этой задаче, выборки X, Y принадлежат нормальной генеральной совокупности с неизвестными параметрами (μ, σ). Для простоты практических вычислений, принимается

аксиоматично следующее утверждение (формулы 11 и 12):

$$(m-1)s_1^2 = \sum_{i=1}^{m}(x_i - \overline{x})^2 \approx \sum_{i=1}^{m}(x_i^2 - m\overline{x}^2) \tag{11}$$

$$(n-1)s_2^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 \approx \sum_{i=1}^{n}(y_i^2 - n\overline{y}^2) \tag{12}$$

$$s^2 = \left[(m-1)s_1^2 + (n-1)s_2^2\right]/(m+n-2) \tag{13}$$

Формула 13 — результирующая дисперсия.

$$(20+10-2)s^2 = (19 \cdot 5.62) + (9 \cdot 2.19) \Rightarrow \hat{s}^2 = 4.52$$

$$\{A\} = 2.49$$

$$\{B\} = 1.13$$

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

s_1^2 можно рассматривать как внутригрупповую дисперсию выборки 1.

s_2^2 — соответственно, как внутригрупповую дисперсию выборки 2.

s^2 — как результирующую дисперсию.

Разность $(s^2 - s_1^2 - s_2^2)$ — как дисперсию, обусловленную совместным влиянием факторов.

$$s^2\left(\frac{1}{m-1} + \frac{1}{n-1}\right) = \frac{s^2}{m-1} + \frac{s^2}{n-1}$$
$$\frac{m+n-2}{(m-1)(n-1)}(\overline{x} - \overline{y}) \tag{14}$$

По формуле 14 определяется различие между выборками. По 14 определяется меж групповая дисперсия. При этом, внутригрупповая дисперсия определяется как сумма величин $\{A\}$ и $\{B\}$.

Если величина 14 не значительна, 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, будет выглядеть как двойная:

$$\sum_s \sum_r (x_{rs} - x_{..})^2 = \sum_s \sum_r (x_{rs} - x_{.s})^2 + \sum_s n_s (x_{.s} - x)^2$$

2012: 1614150 чел./день.

2013: 1599726 чел./день.

(-14424 чел./день).

Стоимость проезда: $\approx$ 22 кроны.

11 Ранговые коэффициенты корреляции Спирмана и Пирсона.

Три варианта проверки:

1. Коэффициент корреляции Пирсона.
2. Ранговый коэффициент корреляции Пирсона.
3. Ранговый коэффициент Спирмана.

11.1 Коэффициент корреляции Пирсона

Даны случайные величины x,y:

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}}, \quad 0 \leq \rho \leq 1$$

Коэффициент корреляции используется при наличии линейной зависимости между случайными величинами. При нелинейной зависимости не применяется. Используется предположение, что выборки x,y одинакового объёма.

11.2 Ранговый коэффициент корреляции Пирсона

Пример: Дано: данные по защите дипломных работ в ВУЗе с 2003 по 2007. Указывается процент отличных оценок, хороших и удовлетворительных и общее количество выпускников. Определить существует ли зависимость между количеством выпускников и распределением оценок на защите дипломных проектов.

	"Отл %	"Хор %	"Удовл %	Общее кол-во выпускников, чел
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

Признак квалификации (реальная метрика, используемая министерством образования):

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

Расчёт статистики $\hat{\chi}^2$ (часть чисел придумана).

	"Отл. %	"Хор. %	"Удов. %	Общее кол-во выпускников, чел.
2003	$\frac{(1485 - 1485 * 0.8)^2}{1485} \simeq 0.1$	...		
	$\vdots$	$\ddots$		
$\sum$	3.85	2.95	0.60	
	7.4			

В рамках решения задачи, проверяется гипотеза о зависимости признаков (нуль-гипотеза), альтернативной является гипотеза о независимости. Для проверки гипотезы используется статистика $\hat{\chi}^2$. По таблицам распределений, при заданном уровне значимости (к примеру 0.01).

Значение 7.4 попало в основную область.

Вывод: Расчитанное значение статистики находится в основной области, т.е. признак квалификации и общее количество выпускников являются зависимыми. Количество выпускников за пять 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.

Величина p может вычисляться по-разному, в зависимости от алгоритма формирования признака квалификации.

11.3 Ранговый коэффициент Спирмана.

Пример: Данные по защите дипломного проекта группы ВУЗа и данные по оценкам абитуриентов - средний балл на вступительных экзаменах.

Ранг за дипломное проектирование (баллы → ранг): 2 → 1, 3 → 2, 4 → 3, 5 → 4. Ранг а вступительный экзамен(баллы → ранг): <3,25 → 1, 3.25-3.50 → 2, 3.75-4.25 → 3, 4.5-5 → 4.

	Д.П. баллы	Д.П. ранг	В.Э. баллы	В.Э. ранг	Квадрат разницы рангов
1	5	4	4.25	3	1
2	5	4	4.50	4	0
3	4	3	3.25	2	1
4	5	4	4.25	3	1
5	4	3	4.50	4	1
6	4	3	4.00	3	0
7	3	2	3.00	1	1
8	4	3	3.50	2	1
9	4	3	3.50	2	1
10	2	1	3.25	2	1
11	4	3	4.00	3	0
12	4	3	3.50	2	1
13	5	4	4.50	4	0
14	5	4	4.75	4	0
15	3	2	3.00	1	1
16	5	4	4.25	3	1
17	5	4	4.75	4	0
18	5	4	5.00	4	0
19	4	3	4.00	3	0
20	3	2	3.25	2	0
$\sum$					10

$(R_{gu/7/|oM\ 6a/|/|b|} - R_{BcTyp.\ eK3})^2 = 10$

В рамках задачи проверяется гипотеза независимости признаков (нуль-гипотеза). Аль-

тернативная гипотеза - функции зависимы. Выбирается статистика (статистика Спирмана):

$$\rho_s = 1 - \frac{6}{n(n^2-1)} * \sum_{i=1}^{n} (R_i - S_i)^2 \simeq 0.9$$

Спирман показал, что величина $\sqrt{n-1} * \rho_s \sim N(0,1)$ (распределена по стандартному нормальному закону).

Задаёмся уровнем значимости по статистической таблице определяем размер критической области (уровень значимости 0.05)

Вывод: Гипотеза о зависимости признаков отвергается и принимается гипотеза что признаки независимы [2].

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

Рис. 8 – Пражские самые поездатые поезда

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

x_1 - число вошедших в течение дня,

x_2 - число купивших билет (по категориям),

x_3 - число прошедших через турникеты.

[2]Подробности в паблике "Типичный пономарь": https://vk.com/tipichnyeponomari

Метро воспользовались за день в 2012: 1.614150 млн. человек. В 2013 году повышают стоимость поездки с 16 до 22 крон. Метро воспользовались за день в 2013: 1.599726 млн. человек.

$$f(x_2, x_3) = \hat{S}(x_2) \cdot \frac{x_1}{\hat{k}}$$

$$x_1^{(6)} = kx_2^{(6)} \to \hat{k} - ?$$

$$x_1 = k * x_2$$

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

По известному значению $x_3 = \frac{x_1}{\hat{k}} = \frac{1.599726}{\hat{k}}$.

Получим, что $x_1 = 1.117 \cdot x_3^{0.975} \approx 1.121041$

Вектор $\{P_D, P_F, \} = \{0.5, 0.24, 0.24\}$

Доверительный интервал для x_3 имеет вид: $[21.772, 41.366]$, $\gamma = 0.999$

Выводы:

- 4/5 пользователей метрополитена являются безбилетниками; С вероятностью $\gamma = 0.999$, оплаченные поездки, или реальный доход оценивается величиной $x_3 \in [21.772, 41.366]$ (реальный дневной доход в 2013 году, в отличие от года раньше, составил 25 млн. крон);

- Можно уменьшить величину доверительного интервала, засчет оценки стоимости долговременных билетов (проездных, трехдневных, и льготных категорий (долговременные));

- Доверительная вероятность оценивается по диапазону $[21.772, 41.366]$;

- Выявлена зависимость 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 расчетный период.

12 Факторный анализ.

12.1 Модель факторного анализа.

$$y = \alpha_0 + \alpha_1 x_1 + \ldots + \lambda_n x_n$$

Позволяют выразить x_i в следующем виде

$$\begin{cases} x_i = \beta_0 + \ldots + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} + \ldots + \beta_n x_n + \epsilon_i, i = \overline{1,n} \\ \sigma_i = \sigma(\epsilon_i) \end{cases} \tag{15}$$

Полагается, что значение каждого признака x_i могут быть выражены как взвешенные суммы других признаков (латентные факторы/признаки/переменные), количество которых может быть меньше чем число исходных признаков.

В 15 ϵ_i — не специфический фактор, а остаточный член ряда, определяющий влияние неучтенных факторов. Можно говорить, что дисперсия x_i фактора, зависит от этого остаточного члена ряда $\sigma_i(\epsilon_i)$

Коэффициенты при переменных называются нагрузкой фактора. Переменные x_i — факторными. Величины $\epsilon_i, i = \overline{1,k}$ независимы друг от друга, и от любого фактора x_i. Можно наложить условия для n признаков, оптимальное число факторов k определяется эмпирической формулой:

$$(n-k) < \frac{(n+k)}{2}$$

Оставим 4 фактора из 10 признаков.

Сумма квадратов нагрузок основной модели анализа называют *общностью соответствующего признака* x_i . Чем больше это значение, тем лучше описывается решение задачи выделенным фактором. Общность есть часть дисперсии признака, котрую объясняют интерпретируемые факторы. В свою очередь, ϵ_i показыает, какой вклад внесли неинтерпретируемые факторы. В связи с этим, общность называют характеристикой специфичных признаков в качестве альтернативы неспецифичного x_i. Основное показывает, что коэффициент корелляции двух любых признаков можно вычислить суммой произведения нагрузок некорелляруемых факторов.

Формально, выражение 15 имеет k неизвестных, и k уравнений, т.е. задача распределения нагрузки решается однозначно. С другой стороны, наложение доп. условий может привести к увеличению неизвестных и неоднозначному решению. В этом случае, можно уменьшить количество факторов засчет вращения системы в заданной системе координат.

Во избежание снижения точности, осуществляют поворот гиперсферы данных в пространстве, при переходе к новой, ортогональной системе координат. В этом случае, разброс ряда переменных в новой системе координат, оказывается не значительным, в силу чего изменением этой координаты можно принебречь. Такой подход называется *вращением факторов*, и лежит в основе *метода главных компонентов*.

В предположении, что наборы коэффициенов β_j для каждого g отличается, часто применяют *сингулярный анализ*:

$$B = \begin{pmatrix} \beta_1^1 & \beta_2^1 & \dots & \beta_k^1 \\ \beta_1^2 & \beta_2^2 & \dots & \beta_k^2 \\ \beta_1^3 & \beta_2^3 & \dots & \beta_k^3 \\ \dots & \dots & \dots & \dots \\ \beta_1^k & \beta_2^k & \dots & \beta_k^k \end{pmatrix} = USV^T$$

На главной диагонали расположены сингулярные числа. Признаки, соответствующие минимальному сингулярному числу, имеют минимальный разброс, а следовательно на результат решения задачи. Это преобразование является аналогом афинного преобразования, и что мы осуществляем переход к новому афинному базису. Факторный анализ реалиован во всех пакетах символьных вычислений. Процедуры работают однотипно: начиная с однофакторного анализа, затем проверяется насколько корреляционная матрица, восстановленная по однофакторной модели, отличается от корреляционной матрицы исходных данных. На каждом этапе проверяется сответствие корреляционных матриц до тех пор, пока не будет выбрано оптимальное количество параметров, либо (что тоже бывает при автоматизированном использовании ФА) совпадет с n.

Вращения факторов может производиться разными способами, что зачастую приводит к неподдающимся содержательной интерпретации факторов. После выбора основных факторов, вращение продолжают до тех пор, пока факторы не окажутся поддрающимися интерпретации. Можно, например, вращать таким образом, чтобы "исчезли трудноинтерпре-

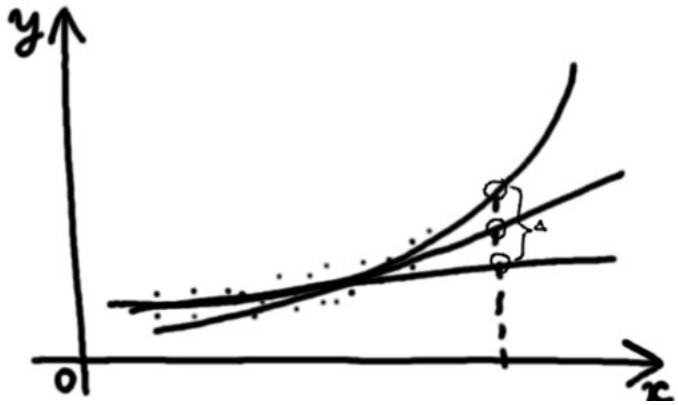
тируемые нагрузки".

Основным недостатком метода является неоднозначность полученного решения. В ряде предметных областей, факторный анализ — превычное и единственный инструмент.

Для выбора решения проводят анализ вклада пары факторов методом дисперсионного анализа. Если ДА и ФА дают сходные результаты, то решение можно считать однозначным.

13 Цензурирование выборок, и анализ выбросов.

У нас есть данные, которые нужно аппроксимировать с помощью метода наименьших квадратов. Нужно разделять два понятия: цензурирование выборок, и анализ выборов.



Цензурирование выборки — замена реальных значений случайных величин выбранными анчёалогами.

Анализ выбросов — изъятие значения(ий) случайных величин из результатов эксперимента на основе статистического анализа данных.

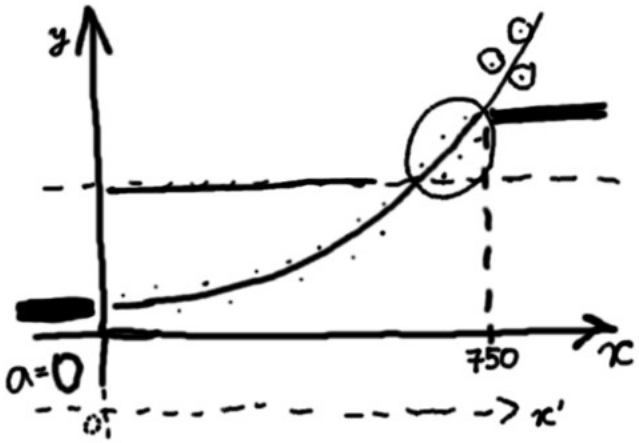


Рис. 9

В случае, если предполагаемые реальные значения близки к заданному значению левой границы выборки a, но не могут быть измерены, то реальные значения заменяются на величину a (*цензурирование слева*) (см. рис. 9) .

Если предполагаемые значения близки к правой заданной границе выборки b, но не могут быть измерены, то реальные значения заменяются на величину 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 и справа попадает не более 10% элементов выборки. В противном случае рекомендуется изменить условия проведения эксперимента. + обеспечиваем точность анализа - если выбрали неправильно - снижаем точность анализа

Анализ выбросов является процедурой альтернативной цензурированию и позволяет выявить точки, искажающие решение задачи.

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

- Базирующийся *в виде закона распределения*;
- Базирующийся *на построении доверительного интервала.*

К группе A относятся метод, основанный на предварительном знании вида функции распределения исследуемой случайной величины, наблюдаемой в рамках эксперимента.

В курсе лекций был рассмотрен пример о проверке (лекция 6) принадлежности выборки генеральной совокупности, распределённой нормально с выбранными параметрами μ и σ.

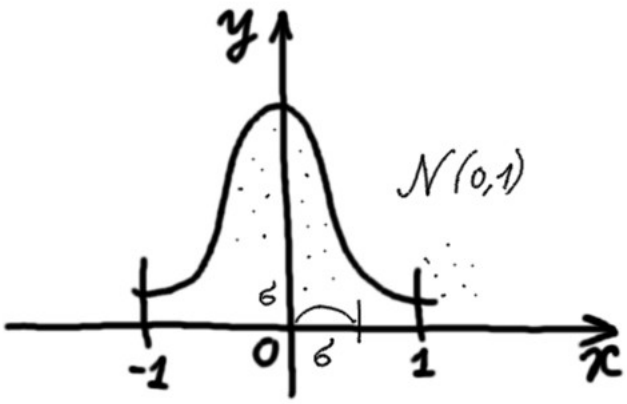


Рис. 10 – Среднеквадратическое отклонение σ нормального распределения $N(0,1)$.

Если закон распределения выборки $x \in X : N_X(\mu, \sigma)$ то анализ выбросов проводится на основе правила 3σ). Процедур, относящихся к классу A немного, так как они специфичны для конкретного закона распределения, кроме того, закон распределения может быть неизвестен.

Альтернативой, обеспечивающей анализ выбросов, является подход, связанный с построением доверительных интервалов, если значительная часть элементов выборки оказыва-

ется вне интервала, то говорят о наличии выбросов.

$$\left[\overline{x} - k\hat{S}, \overline{x} + k\hat{S}\right], \quad \gamma = 0.99$$

(На экзамене требуется воспроизвести только структуру полученного доверительного интервала.)

Существует достаточно много вариантов оценок коэффициента k, в ряде случаев анализируют выборочную функцию распределения, и фиксируют квантили заданного уровня. Если при проведении другого эксперимента большее количество элементов выборки, чем в контрольной, оказывается вне диапазона $[k(\alpha_1), k(\alpha_2)]$ говорят о наличии выброса в выборке.

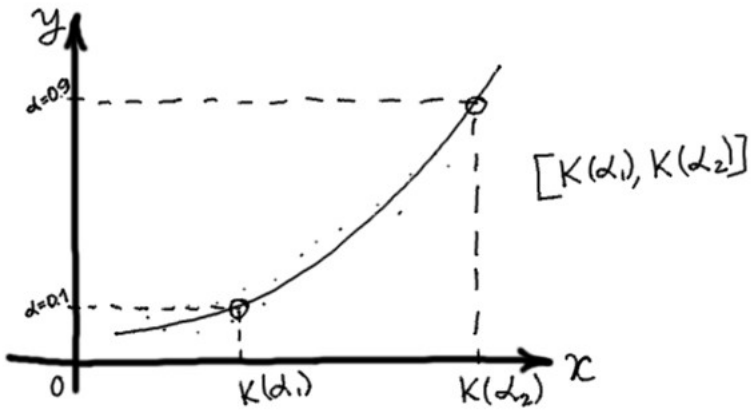


Рис. 11 – Квантили уровней α.

Недостатком методов группы B является выявление значительного количества выбросов выборке (x, y), где x – число вошедших в метро. Всего будет 12 ф-ций (для входящих на станцию, и для прошедших через турникет). Строим 12 выборочных функций распределений.

14 Моделирование на ЭВМ случайных величин, векторов, процессов.

В ряде слчаев возникает необходимость моделирования данных на ЭВМ. В этом случае говорят, о примененении *генератора псевдослучайных чисел* (ГПСЧ) по заданному закону распределения.

Различают три вида генераторов случайных чисел:

1. **Физические ГСЧ.**

   - бросание монеты;
   - физические приборы и датчики;
   - таймеры.

   Достоинства:

   - Точность моделирования

   Недостатки:

   - Ограниченный ряд законов распределения;
   - Неудобство использования.

2. **Табличные ГСЧ.** Достоинства:

   - Высокая точность моделирования.
   - Доказана принадлежность к определённому закону распределения.

   Недостатки:

   - Большой объём справочной информации.

3. **Математические ГСЧ (ГПСЧ).** Первый (арифметический) гпсч был предложен Фон-Нейманом имел очень малый период (быстро возникало зацикливание либо себя, либо 0). Модификации арифметического способа предполагали осуществление сдвига влево или вправо, что позволяло увеличить период метода (до 1000 значений без зацикливания). Тем не менее, подход был малопригоден

для решения задач на компьютере. В настоящее время используется линейный конгруэнтный подход.

Строится рекуррентная последовательность:

$$X_{n+1} = (aX_n + b) \mod m$$

Подход предложил Лимер (1949г). Предложенный генератор предполагает выбор трех чисел a, b, m , и начальное значение x_0 — берётся с таймера.

mod m — операция взятия по модулю m При удачном выборе начальных чисел генерируемая последовательность содержит независимые случайные величины.

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

- *Метод, на основе центральной предельной теоремы (ЦПТ).* Известно, что сумма нескольких независимых с.в., равномерно распределенных в интервале $(0, 1)$ асимптотически стремится к нормальному распределению, т.е. имеет асимптотически нормальное распределение.

$$r_i : R\,[0, 1]$$

$$x = \sum_{i=1}^{n} r_i, \quad x \in N(\mu, \sigma) \quad \mu = 3.5, \ \sigma \approx 1.7$$

$$[2, 1, 5, 3, 4, 6]$$

$$\frac{1}{6}(1.5^2 + 2.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 2.5^2) = \frac{1}{6}(4.5 + 0.5 + 12.5) = 17.5/6 \approx 2.91$$

- *Метод обращения.* Предполагается, что плотности распределения ставится экс-

поненциальный закон распределения.

$$F(x)\tilde{}f(x)$$
$$F(x) = 1 - e^{-\lambda x} \Rightarrow \lambda = \frac{1}{x} \tag{16}$$
$$f(x) = \lambda e^{-\lambda x}$$

Рассчитывается выборочное среднее и строится выборочная функция распределения, которая может быть описана аналитической фукцией вида 16. Строится функция, обратная функции распределения, которая выражается через функцию плотности распределения с неизвестным параметром λ, откуда может быть определён параметр λ.

$$F'(x) = f(x) = \frac{1}{\lambda} f'(x) \to 0 + \frac{f(x)}{f'(x)} = \frac{1}{\lambda} f^{-1}(x) \Rightarrow x = 1 - F(x) = e^{-\lambda x} = \frac{f(x)}{\lambda} \tag{17}$$

На основе предположения $f(x_0) = 0$, то ... в левой части оказывается уравнение касательной, рассматриваемая как сумма приращений аргумента x

Метод обращений реализуем для узкого круга функций, и как правило используется для экспоненциального закона. На этом калитка и закрывается.[3]

- *Метод Неймана.* Подход является универсальным для любого метода распределения

[3] "А мне уже не спится, и в моменте кое-что вспоминается ..."©

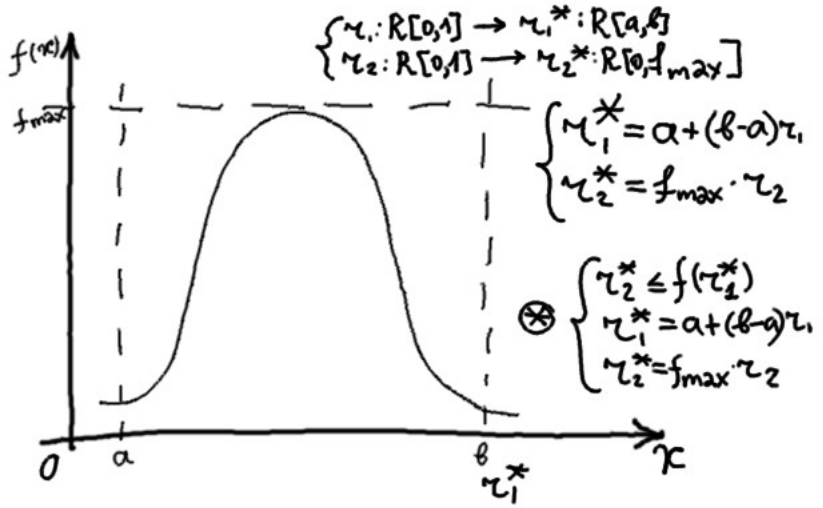


Рис. 12

Если пары точек удовлетворяют условию ограничения системы (*), то, соответствующие им значения r_1^* составялют выборку из генеральной совокупности, распределённой по заданному закону, в противном случае отбрасывается. Метод, конечно, очень хороший

Если распределение имеет сложный вид, то его декомпозируют на простые по форме составлящие, интерполируют "простыми"функциями распределения (чаще равномерным). Такой метод называется *методом суперпозиций*.

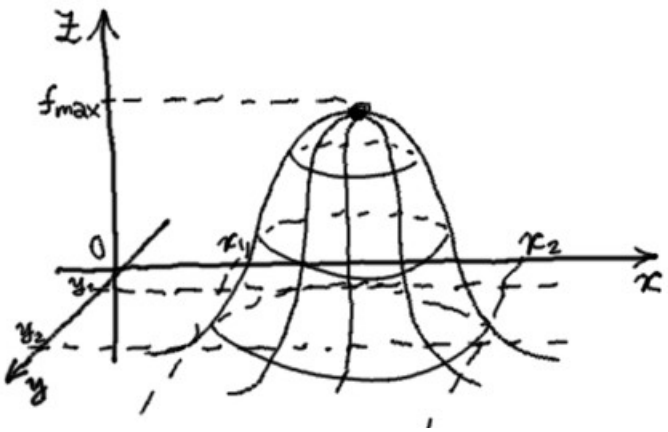
В предположении, что случайный вектор представляет собой набор зависимых случайных величин с заданным законом распределения, моделирование случайного вектора не отличается от подхода к моделированию случайных процессов, так как случайные процессы (функции) на ЦВМ дискретизируются.

Рассмотрим моделирование случайного процесса с показательным распределением. В этом случае получают последовательность n независимых случайных величин (с нормальным стандартным распределением), после чего строится рекуррентная зависимость:

$$\xi(t_n) = \rho_n \xi(t_{n-1}) + \sqrt{1-\rho_n^2} x\,[n]\,,$$
$$\rho_n = e^{-\lambda(t_n - t_{n-1})}, \quad \lambda = \frac{\pi}{2\Delta t}$$

Показательное распределение (экспоненциальный закон) является примером моделирования стационарного случайного процесса с заданным законом распределения. Другие законы распределения предполагают более сложную схему описания, в силу чего в ряде случаев задачи моделируются процессы в иной классификации (хотим абстрагироваться от закона распределения). Универсальным подходом к моделированию стационарного случайного процесса с заданным законом распределения является обобщение метода Фон-Неймана на n-мерный вектор.

Формируется набор из $n+1$ случайной величины, распределенной равномерно. Проводится переход преобразования:



$$r_1 \to r_1^*$$

$$r_2 \to r_2^*$$

$$r_n \to r_n^*$$

$$r_{n+1}^* = f_{max} r_{n+1}$$

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

где значения со звёздочкой изменяются в любом заданном диапазоне не превышает f_{max}. Если выполняется условие $**$ набор сохраняется и координата r_{n+1} считается распределённой по определённому закону в диапазоне $[0, 1]$. В противном случае он отбрасывается.

Недостатками являются:

- Усечение функции плотности распределения;
- Получение набора независимых случайных величин.

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

Марковский случайный процесс — 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 понятия *марковской цепи* – частный случай марковского процесса, когда состояние процесса дискретно. Кроме того марковский процесс рассматривают как авторегрессию первого порядка

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

Определение марковского процесса по Венциль — «будущее» процесса зависит от «прошлого» только через его «настоящее».

Пример: Известны вероятности перехода из одного состояния в другое (s_0 — новая мишень, s_1 — повреждёння мишень, s_2 — поражённая мишень). Необходимо определить 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}	
S_2	0	0	P_{22}	

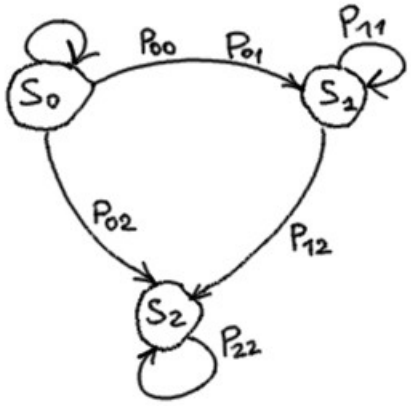


Рис. 13 – Граф на основе таблицы

$$1) \quad S_0 \to s_0 \to S_1 \to s_1 \to s_1 \to s_2 : 5$$

$$2) \quad s_0 \to s_0 \to s_0 \to s_2 : 3$$

$$\dots$$

$$8) \quad s_0 \to s_2 : 1$$

При увеличении числа реализаций случайного процесса оцениваемая величина стремится к величине a и при бесконечном числе реализации достигает его. Для остановки счёта на ЭВМ пользуются следующим правилом:

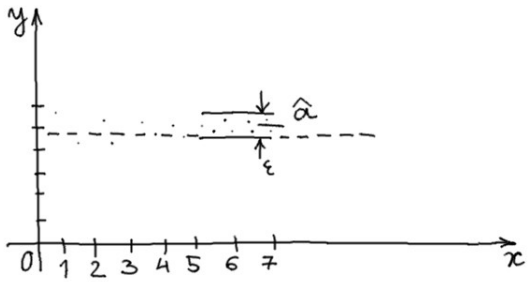


Рис. 14

Если 3 последние точки попадают в заданный диапазон с разницей границ, не превышающей , то счёт завершается, а оценкой a является середина заданного диапазона. (см. рис. 14). В данном случае величина равна 4.25 (диапазон [4, 4.5]). Т. к. речь идёт о ресурсах, то округление в большую сторону.

Ответ: в среднем потребуется 5 снарядов.

15 Классификация современных средств моделирования на примере пакета MathWorks.

Решение компании MathWorks представляют современный инструментарий проектирования и имитации сложных систем на основе заданной или построеной математической модели, позволяет создавать сложные *многодоменные* объекты, включающие приложения для обработки изображений сигналов, математических расчетов, расчетов технической оснаски проекта, обмена данными, и т.д.

Многодоменные — подразумевается имитация асцилографиа, компьютера, и для каждого оборудования производится имитация.

Mathworks предполагает описание всех аппаратных, программных и алгоритмических средств участвующих в моделировании — как объектов, характеризуя их наборами данных, в т.ч. моделями, ф-циями и возникающими, в процессе функционирования системы, событиями. Такой подход называется *объектно ориентированное проектирование*.

15.1 Построение моделей сложных систем.

Наследуя принципы среды моделирования Simulink, основанной на графических блок-схемах, MathWorks позволяет выбрать структуру описываемой системы, и установить связи разного типа между компонентами системы.

Вводится набор дескрипторов, определяющий каждую компоненту как объект, т.е. каждый блок может быть описан физически, математиескими алгоритмами, дискретными событиями или графиками состояний.

Оптимизация схемы (удаление, объединение) блоков функционирование которой не влияет на решение задачи.

Выбирается способ имитации системы, реализуются алгоритмы посредством библиотеки мат. алгоритмов, унаследованные от Matlab и Simulink.

Достоинства пакета Mathworks:

- Интеграция большого объема вычислительных и аналитических методов;
- Удобный графический интерфейс;

- Интерфейс с другими языками программирования, в т.ч. импорт/экспорт программ C/C++, HDL;
- Возможна автоматическая генерация кода на C/C++; Код генерируется непосредственно из модели сложной системы. Удобно для развертывания и прототипирования собственной системы;
- Встроенный язык для распараллеливания вычислений, в т.ч. распараллеливание непрерывно постуающих с датчиков данных;
- Встроенная система тестирования.

Недостатки пакета:

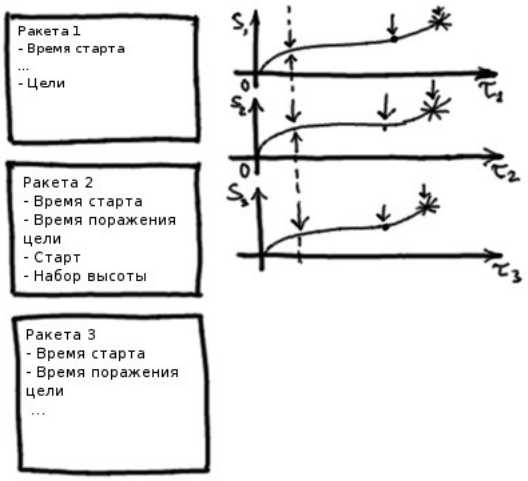
- Точность вычислений;
- Значительное количество ошибок возникает при распараллеливании вычислений и данных;
- Ограничен интерфейс с иными средствами программирования.

16 Объектно ориентированное моделирование.

Объектно ориентированное моделирование позволяет оптимизировать, а также верифицировать, обеспечить валидацию требовний, и осуществить тестирование модели.

16.1 Верификация и валидация тестирования.

Целью процесса является выявление ошибок на ранних этапах моделирования. Для каждой компоненты системы вводится так называемая спецификация, включающая, помимо атрибутов и операций компоненты, наборы входных/выходных данных в заданный или формируемый момент времени.



Верификация выполняется на всех этапах функционирования системы для каждой компоненты. Верификация должна учитывать взаимное влияние компонент и изменение условий моделирования. Процесс оказывается трудоёмким 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 противоречивые). Возникает необходимость **ранней валидации требований**. Как правило модель содержит ограничения, накладываемые на переменные с учётом физической природы изучаемой величины, изменений окружающей среды или условий проведения эксперимента, а также накоплением вычислительной погрешности при реализации модели. В случае, если эти составляющие на практике учесть сложно, то исследователь вводит доверительный интервал с заданным коэффециентом доверия для исследуемой величины. В ряде случаев исследователь затрудняется в оценке границ доверительного интервала с высоким

коэффециентом 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 для проверки системы. При реализации модели, величины, исследуемые на таких пространствах, должны попадать в заданные диапазоны. Для справки: В MathWorks Реализован где-то там модуль `SystemTest`, который на основе продукта Simulink Verification Validation. Разработчик может связать свою схему модели и диапазоны изменения величин со стандартными процедурами и тестами этих продуктов.

17 Математические основы теории массового обслуживания (ТМО).

Создателем ТМО считается советский математик А.Я. Хинчин. Стартом для его исследований явились работы математиков английской актуарной школы и копенгагинской телефонной компании (в частности А. Эрланг). Пусть имеется телефонный узел (устройство, прибор в терминологии ТМО), на котором телефонистки соединяют пары телефонных абонентов. При небольшом количестве звонков соединение не требует ожидания. При интенсивном увеличении говорят об *СМО с ожиданием*.

Ожидающие удовлетворения заявки (транзакты) помещаются в очередь. Очередь может быть ограничена (N заявок). В этом случае говорят о возможности потери заявок (всего L заявок). Если считать заявки (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 способа действия*.

$$[t, t + \Delta t] \tag{18}$$
$$[t^*, t^* + \Delta t]$$

Поток без способа действия имеет действие если количество обработанных заявок в любом временном интервале остаётся постоянным в любом совпадающем по длительности непересекающемся с исходным интервале времени. Поток заявок *стационарен* — если вероятность обработки n заявок в интервале (см. формулу 18) не зависит от времени t, а зависит только от Δt (длительности интервала). Однородный стационарный поток без способа действия называется простейшим потоком Пуассона. Число событий такого потока распределено по закону Пуассона.

Мгновенная плотность потока — предел отношения среднего числа заявок обработанный в элементарный интервал $[t, t + \Delta t]$ к длине этого интервала, при $\Delta t \to 0$. В технических приложениях называется *интенсивностью потока*.

Для простейшего потока:

$$\lambda = \frac{M(t)}{\Delta t}$$

Среднее количество заявок в системе определяется формулой:

$$N = \lambda T \tag{19}$$

T – время обработки;

λ — интенсивность;

N — среднее количество заявок в системе (это интенсивность на время обработки).

Формула 19 называется *формулой Литтла* и позволяет оценить среднеее количество заявок в системе. Основными элементами СМО (помимо входного потока заявок, очереди заявок) являются каналы (несколько однотипных приборов обслуживания), выходной поток обслуженных заявок, фаза обслуживания. Системы делятся на:

- одноканальные;
- многоканальные.

и на системы:

- С очередями ожиданием;
- С очередями отказов;

По типу равноценности заявок на системы:

- С приоритетом;
- Без приоритета.

по фазам обслуживания:

- Однофазные;
- Многофазные.

На определённых фазах возможно повторное обслуживание одной и той же заявки.
По взаимосвязи с потоками заявок, системы делятся на:

- открытые (разомкнутые);
- замкнутые.

Если интенсивность входного потока заявок не зависит ни от количества заявок в СМО, ни от количества уже обслуженных заявок, то говорят об *открытой СМО*. Система, сочетающая в себе свойства многоканальности, многофазности, разомкнутости классифицируется как **сеть** массового обслуживани.

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

- **Абсолютная пропускная способность СМО**: среднее количество заявок, обслуживаемых системой в единицу времени.
- **Относительная пропускная способность СМО**: отношение среднего количества заявок, обслуживаемых системой в единицу времени к среднему числу всех заявок, поступивших за это время в СМО.

Среднее число занятых каналов и коэффициента их занятости (отношение числа занятых каналов к общему числу каналов) формализуются как показатель эффективности СМО. Среднее число свободных каналов и коэффициент простоя (формализуются по паказателям занятости (использования) каналов). Среднее время нахождения заявки в очереди, среднее врем нахождения заявки в СМО, а также дисперсия числа заявок в очереди и в целом, в СМО, также являются показателями эффективности СМО. Указанные показатели легко формализуются и являются аналитическими моделями СМО.

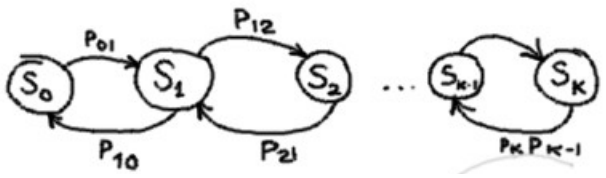
17.1 Уравнение Колмогорова.

Состояние СМО определяется количеством занятых каналов обслуживания и числом мест в очереди. Очевидно, что 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 марковская цепь дискретна, те 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, то вероятности для марковских цепей могут быть рассчитаны

посредством разностных схем. А в случае непрерывных марковских систем Du колмогорова (Дифференциальным уравнением)

18 Процессы гибели-размножения.

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



$\sum p_{ij} = 1$ (для состояния i)

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

$$\begin{cases} S_0 \to S_1 \\ \lambda_{01}p_{01} = \lambda_{10}p_{10} \\ \cancel{\lambda_{01}p_{01}} + \lambda_{12}p_{12} = \cancel{\lambda_{10}p_{10}} + \lambda_{21}p_{21} \\ \cancel{\lambda_{01}p_{01}} + \cancel{\lambda_{12}p_{12}} + \lambda_{23}p_{23} = \cancel{\lambda_{10}p_{10}} + \cancel{\lambda_{21}p_{21}} + \lambda_{32}p_{32} \\ \dots \\ \lambda_{ij}p_{ij} = \lambda_{ji}p_{ji} \\ \dots \end{cases}$$

Мы можем последовательно сначала сократить элементы 01 и 10, затем 12 и 21 и т.д. В итоге получим систему:

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

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

КТО ВИНОВАТ? И ЧТО ДЕЛАТЬ? Что мне тогда отвечать:

Неизвестных в два раза больше чем уравнений. Для сокращения числа неизвестных в

2 раза предполагают равновероятными переходы из одного и того же состояния S_1 в предыдущее и следующее. То есть равновероятным является переход в новое состояние на 1 шаг и возврат в предыдущее.

$$(**)\begin{cases}\lambda_{01}p_1 = \lambda_{10}p_2\\ \lambda_{12}p_2 = \lambda_{21}p_3\\ \lambda_{23}p_3 = \lambda_{32}p_4\\ \dots\\ \lambda_{k-1,k}p_k = \lambda_{k,k-1}p_1\end{cases}$$

Система (**) называется системой уравнений гибели-размножения. Решается на изи вообще. Все вероятности выражаются через вероятность p_1. Из последнего уравнения (**) выразим p_1.

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

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

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 функции марковского случайного процесса описываются как дифференциальные уравнения Колмагорова, в предположении, что функция перехода из состояния i в состояние j зависит от моментов перехода s и t.

$$\frac{dp_{ij}(s,t)}{ds} = \sum_k \alpha_{jk}(s) p_{kj}(s,t), \quad i = \overline{1,k}, j = \overline{1,k}$$
$$\frac{dp_{ij}(s,t)}{ds} = \sum_k p_{ik}(s,t) \alpha_{k,j}(t), \quad i = \overline{1,k}, j = \overline{1,k}$$
$$\alpha_{ij}(s) = \lim_{t \to s} \left[p_{ij}(s,t)_{1-\sigma_{ij}} \right] / (t-s), \quad t > s.$$
$$1 - \sigma_{i,j} \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Можно сформулировать правила составления уравнений Колмогорова по размеченному графу состояний непрерывной марковской цепи.

1. Число уравнений в системе **равно** числу вершин графа;
2. Вероятность состояния p_i соответствует как переходу в последующее, так и возврату в предыдущее состояние.
3. Система уравнений имеет форму Коши;
4. Число слагаемых в правой части равно числу дуг в графе, интерпретирующих переход из i-го состояния в любое другое, кроме самого себя;
5. Переход в новое состояние соответствует слагаемое со знаком «+»;
6. Возврату в предыдущее состояние соответствует слагаемое со знаком «-»;
7. Каждое слагаемое представляет собой произведение вероятности i-го состояния и плотности вероятности перехода по данной дуге;
8. Начальные условия для постановки задачи Коши определяются непосредственно начальным состоянием системы; Например, в цепочке $s_0, s_1, \ldots s_k$ старт начинается из состояния s_2;

$$p_0(0) = 0; \;\; p_1(0) = 0; \;\; p_2(0) = 1, \ldots, p_k(0) = 0.$$

19 Подготовка к экзамену

Распределение Вейбола: (Синтезированное, искусственное распределение).

Можно дополнительно рассмотреть распределения Эрланга, Стьюдента, и т.д.

Вопрос: Метод Неймана (для случайных величин — пару, а для n — n + 1).

Метод суперпозиции — рассматриваем на каждом отдельном фрагменте разное распределение. Интерполяция набором простых распределений.

19.1 Вопрос

Случайная величина:

$$\sum_i x_i, x_i : N(0,1), i = \overline{1,6}$$

1. Сгенерировали 6 значений (так решили исследователи). Получают $x_i = \sum_{i=1}^{6} r_i$;
2. Повторяем шаг 1) 20 раз x_j, $j = \overline{1,20}$;
3. посчитали выборочное среднее и выборочную дисперсию (СКО) σ_x^2;
4. Перешли к величине $N(0,1)$;
5. Переход к выборке с заданными параметрами.
   $x_j : N(\overline{x}, \sigma_x^2) \rightarrow$
   $x'_j : N(0,1)$;
6. Переход от выборки $x'_j : N(0,1) \rightarrow x''_j : N(\mu, \sigma^2)$.

Потом от выборки $N(0,1)$