

Современные методы анализа данных в задачах управления – СМАДЗУ

6 курс

Лекция 1

Проф. Ососков

Геннадий Алексеевич

Семестровый курс со сдачей зачета

Большие данные окужили нас в повседневной жизни

Компьютеры, смартфоны
GPS-навигаторы
WWW: Google, Яндекс,
вконтакте, одноклассники,
Facebook,
LinkedIn, ResearchGate
TV,
электронные книги

Информация
Данные
Картинки
Аудио
видео



Бурный рост потоков данных за последние годы, особенно в социальных и бизнес приложениях, где изучаемые явления слишком сложны для их математической формализации, вызвало насущную **потребность в выявлении зависимостей напрямую из огромных массивов данных.**

К тому же, только очень малая часть этих данных будет когда-либо востребована, потому что их **объемы слишком велики**, чтобы их вместить в имеющиеся базы данных, а **структура данных слишком сложна или, вообще, не определена** для эффективного их анализа **имеющимися** алгоритмами **за разумное** время.

О пользе и опасности БОЛЬШИХ ДАННЫХ - читайте эту книгу



<http://www.livelib.ru/book/1000755419>

Большие Данные

Определение: Большие Данные - те, что слишком велики и сложны, чтобы их можно было эффективно запомнить, передать и проанализировать стандартными средствами доступных баз данных и иных имеющихся систем хранения, передачи и обработки.

Если, кроме объема, учитывать и другие их характеристики, то для определения Больших Данных применяется правило **«mpex V»**: объем (**V**olume), скорость (**V**elocity), многообразие (**V**ariety), хотя теперь, когда общий поток данных растет экспоненциально, удваиваясь каждый год, начали добавлять новые «**V**», типа **Value** (ценность), **Veracity** (достоверность) и др., что говорит о расплывчатости этого понятия и поэтому привело к угасанию интереса к этой концепции.

В современных условиях данных слишком много, они неоднородны, неполны, неструктурированы и содержат ошибки, а какой-либо рациональной теории для их описания, как правило, нет.

Поэтому происходит **сдвиг парадигмы** их обработки **от классической схемы**:

1. получение **экспериментальных данных**,
2. моделирование на основе **известной теории**,
3. применение **вычислительных средств анализа** данных для проверки модели путем сравнения с экспериментом

- **к новой 4-й парадигме науки**, когда модели, описывающие связи и зависимости создаются непосредственно из самих данных новыми средствами **Data Mining**.

Data Mining для работы с Big Data

Еще в 90-х появилась технология **Data Mining** (*Добыча данных* или *Интеллектуальный анализ данных*), которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей, необходимых для принятия решений в различных сферах человеческой деятельности таких, как ассоциативные правила, деревья решений, кластеры, математические функции.

Задачи data mining – сбор данных, описательные и предсказательные задачи

1. Сбор, хранение данных: сбор, оцифровка, сжатие, Базы Данных

2. Описательные:

Ассоциация - выявление закономерностей между связанными событиями

Кластеризация - группировка объектов, кластерный анализ;
на основе данных об их свойствах (похожести);

Корреляция - установление статистической зависимости непрерывных выходных от входных переменных

3. Предсказательные:

Классификация - отнесение объектов к одному из заранее известных классов

Поиск функциональной зависимости – регрессионный анализ, анализ временных рядов;

Компьютерное моделирование реальных процессов

– методы Монте-Карло

Методы Data Mining

- **Статистические методы:** анализ связей (корреляционный, дисперсионный и регрессионный анализ, факторный анализ, Фурье и вейвлет анализ).

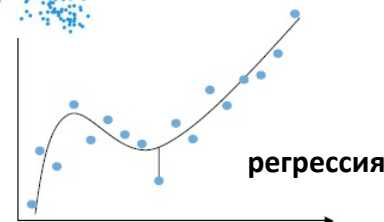
- **Машинное обучение**



оцифровка



кластеризация



регрессия

Машинное обучение

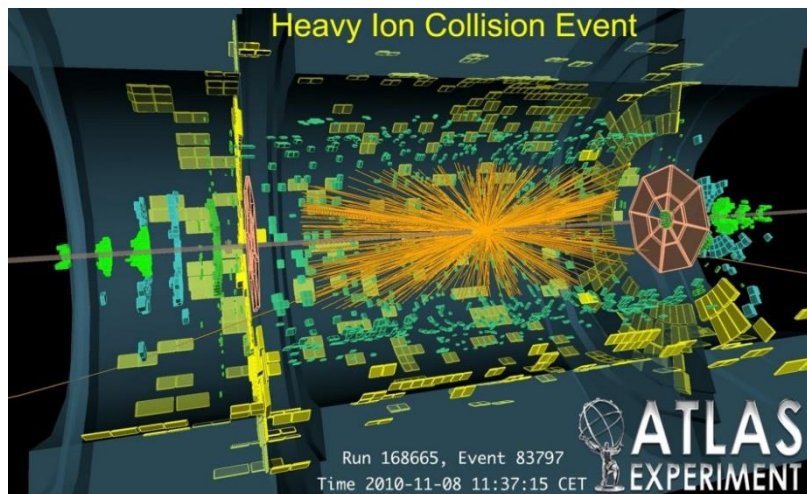
Машинное обучение (Machine Learning-ML), это когда компьютер не просто использует заранее написанный алгоритм, а сам обучается решению поставленной задачи на большой выборке данных.

Три причины синхронного взрыва популярности ML в последние годы:

1. **Большие Данные.** Данных стало так много, что новые подходы были вызваны к жизни тем, что растущее разнообразие, как данных, так и возможных решений стало слишком велико для традиционных заранее запрограммированных систем.
2. Снижение стоимости **параллельных вычислений и памяти компьютеров.**
3. Новые алгоритмы **обучения глубоких искусственных нейронных сетей.**

Большие данные поступают с детекторов экспериментальной физики высоких энергий и ядерной физики

Большой Адронный Коллайдер (Large Hadron Collider - LHC) в ЦЕРНе



Один из 4-х LHC экспериментов - ATLAS



RUN 1
2012
Получен
бозон
Хиггса!

Схема процесса обработки данных эксперимента

1 петабайт=10¹⁵ байт
1 эксабайт=10¹⁸ байт

RUN 2 2015-2017 гг – уже
эксабайты данных в год!

Детекторы LHC-экспериментов выдают данные со скоростью около 1 петабайт/сек

Система интеллектуальных триггеров и фильтров сжимает эти данные в миллионы раз, оставляя на долгое хранение лишь полезную информацию, в итоге **БАК выдает для хранения 50 терабайт в секунду**, - столько данных за 4 часа, сколько вся сеть Facebook собирает за сутки.

Обработать такой объём данных в ЦЕРНе невозможно, поэтому

1. Создана всемирная интернет-сеть распределенных вычислений (**Worldwide LHC Computing Grid -WLCG**)
2. Для моделирования и анализа данных разработаны многочисленные пакеты программ, использующих **методы машинного обучения**

Хранение и обработка данных - ключевая проблема и для других физических центров

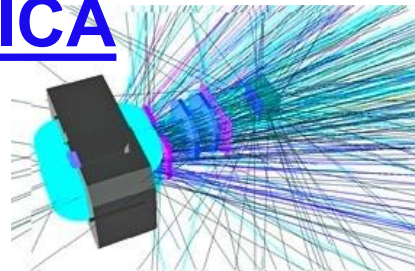
SKA- Square Kilometer Array -

Квадратный километр, занятый радиотелескопами в Южной Африке, будет выдавать ~ **20 экзбайт в год**

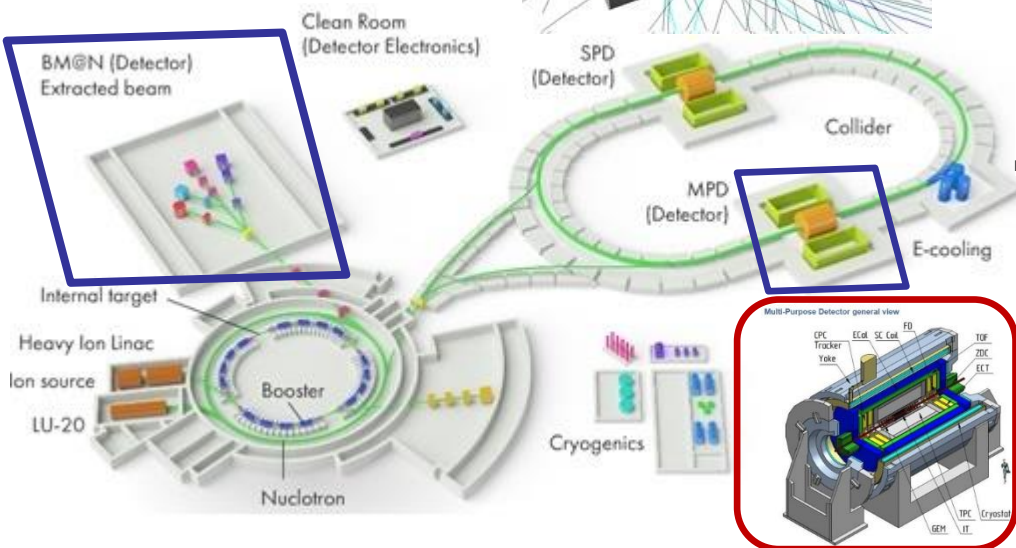
Мегапроект NICA


Эксперимент **BM@N**

Стриповый GEM-детектор внутри магнита



Прогноз производительности NICA
Скорость передачи данных 4,7 ГБ / сек
19 миллиардов событий в год, что после обработки и анализа даёт для хранения - **8.4 PB в год**



Трековый детектор TPC внутри магнита MPD.  Показано смоделированное событие от взаимодействия ионов золота, порождающее тысячи треков

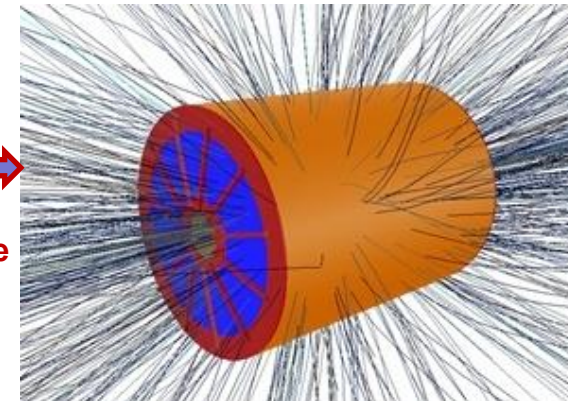


Схема комплекса NICA с экспериментами MPD, SPD, BM@N

Отличия DM и анализа данных в ФВЭ и ЯФ

различие

Физики, захлестываемые потоками данных от экспериментов и моделирования физических процессов, разработали **свой собственный всеохватывающий набор методов анализа данных (Data Analysis – DA)**, реализованный в известной программной платформе **ROOT**, на которой теперь основаны почти все программные оболочки – фреймворки большинства европейских экспериментов. В отличие от DM, методы DA в физике высоких энергий и ядерной физике используют выдающиеся достижения теоретической физики, дающие возможность успешно **моделировать сложнейшие физические процессы, происходящие в экспериментальных установках** при взаимодействиях частиц в каждом из детекторов и траекторий получившихся осколков в каждом из компонентов этих детекторов с учетом их материалов и магнитных полей.

общее

Методы DA – это только часть общего гигантского процесса манипулирования данными в современных экспериментах ФВЭ и ЯФ. Помимо задач анализа данных не менее значительную часть занимают **задачи хранения и обмена данными** в иерархической **ГРИД-облачной системе распределенных вычислений**, объединяющей Tier-центры разных уровней.

Концепция использования распределенных облачных систем для хранения, распределения и обработки данных является общей для физиков и бизнесменов

Этапы процессов DA в ФВЭ и ЯФ 1

Важнейший этап – предобработка включает

- **Получение и сохранение данных:** до применения алгоритмов DA данные, подлежащие исследованию должны быть зарегистрированы, преобразованы из отсчетов детекторов в формат обычных единиц измерений;
- **Селекция данных:** фильтрация от шума и несущественных измерений, не удовлетворяющих заданным условиям. Проверка этих условий выполняется системой «умных» триггеров разных уровней и ведет к сокращению объема данных на много порядков;
- **Преобразование данных (калибровка и алайнмент)** для перевода в формат подходящий для последующего анализа и хранения.

Этапы процессов DA в ФВЭ и ЯФ 2

Следующие этапы можно суммировать как

■ **Распознавание образов для реконструкции событий**: трекинг, нахождение вершин событий, распознавание колец черенковского излучения, а также выявление и удаление ложно распознанных объектов. Применяемые методы

- преобразования Хафа,
- **клеточные автоматы**,
- фильтр Калмана,
- искусственные нейронные сети,
- **вейвлет-анализ** и др.

■ **Оценивание физических параметров**

- **методы математической статистики**;
- **робастное оценивание**

■ **Проверка гипотез**

- Метод отношения правдоподобия,
- **Искусственные нейросети**
- усиленные алгоритмы машинного обучения (boosted decision trees - BDT).

□ **Моделирование выполняется на всех этапах анализа данных**



Методов много и за семестр их не изучить, поэтому успеем только то, что тут выделено

Темы курса СМАДЗУ

(какие задачи будем решать)

- Моделирование случайных воздействий и статистический анализ сигналов
 - Алгоритмы моделирования случайных последовательностей с заданными законами распределения. Их реализация на C++ и в EXCEL
 - Проверка качества моделей по статистическим критериям хи-квадрат и Колмогорова.
- Подгонка зависимостей к данным измерений. Методы максимального правдоподобия (ММП) и Наименьших квадратов (МНК). Робастная подгонка к засоренным измерениям.
- Монте-Карло интегрирование многомерных функций
- Вейвлет-анализ и его применение для обработки сигналов
- Искусственные нейронные сети (ИНС) и клеточные автоматы. Решение задач на их применение

Роль статистики и моделирования

Теоретическая физика разрабатывает сложнейшие аналитические и численные модели как элементарных частиц, так и космических явлений. Эти модели становятся физическими законами только после проверки их соответствия экспериментальным данным.

Однако изучаемые явления

- не наблюдаются непосредственно, а происходят в ходе экспериментов как косвенные проявления взаимодействия ускоренных элементарных частиц с веществом сложнейших детекторов.

- происходят с очень малыми вероятностями $< 10^{-8} - 10^{-9}$

Поэтому для обнаружения столь редких событий требуется "перелопачивать" за короткие сроки многие **миллиарды** наблюдаемых событий, чтобы на **основе их статистического анализа** отбирать только несколько тех, что содержат значимую информацию об изучаемом явлении.

Моделирование позволяет:

- Оптимизировать по деньгам, материалам и времени всю экспериментальную установку и разработать алгоритмы анализа еще на стадии проектирования;
- Разработать и протестировать необходимую программную оболочку эксперимента;
- Оптимизировать структуру и необходимое оборудование запланированных детекторов, минимизируя стоимостные и временные затраты при заданной эффективности и точности работы детектора;
- Рассчитать заранее все необходимые распределения, пороги для проверки гипотез и сгенерировать обучающие выборки для искусственных нейронных сетей.

Принципы моделирования сигналов

- **АНАЛИЗ** - статистический анализ данных об объекте, определение их источника (данные о состоянии элементов объекта управления или данные, получаемые от него самого в процессе его работы), и их природы: детерминированные или стохастические. Для последних – проверка гипотезы об их распределении и оценка параметров этих распределений;
- **СИНТЕЗ** - разработка математической и компьютерной модели, имитирующей как сам объект, так и процесс управления;
- **ВЕРИФИКАЦИЯ и СОПРОВОЖДЕНИЕ** - проверка адекватности модели методами математической статистики, ее отладка и сопровождение.

Напоминание о законах распределения случайных величин

Примеры дискретных распределений

1. Биномиальное распределение. Случайная величина ξ имеет биномиальное распределение с параметрами n и p , если ξ принимает значения $k=1,2,\dots,n$ с вероятностями $P(\xi = k) = C_n^k p^k (1-p)^{n-k}$

Случайная величина с таким распределением имеет смысл **числа успехов в n независимых испытаний** схемы Бернулли с вероятностью успеха p . Среднее значение и дисперсия для биномиального распределения имеют вид:

$M \xi = np$, $D \xi = npq$, где $q=(1-p)$.

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

2. Распределение Пуассона. Случайная величина ξ имеет распределение Пуассона с параметром $\lambda > 0$, если ξ принимает значения $k=1,2,\dots$

с вероятностями

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Математическое ожидание и дисперсия распределения Пуассона равны λ .

Примеры непрерывных распределений

1. Равномерное распределение. Равномерное распределение на отрезке

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases}$$

(a, b) задается плотностью распределения
Математическое ожидание равномерного распределения равно $(b-a)/2$, дисперсия - $(b-a)^2/12$

2. Нормальное распределение. Нормальное распределение с параметрами a и σ задается плотностью.

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

Обозначение: $x \in N(a, \sigma)$

Математическое ожидание нормального распределения равно a , дисперсия - σ^2 .

Стандартное нормальное распределение. Стандартное нормальное распределение - это нормальное распределение с параметрами $a = 0$ и $\sigma = 1$,
Если x имеет стандартное нормальное распределение, то случайная величина $y = x * a + \sigma$ распределена нормально с параметрами a и σ .

3. Показательное распределение задается плотностью
Математическим ожиданием показательного распределения является величина $1/\lambda$, обратная к параметру распределения.
Для $x > 0$ функция распределения равна $F(x) = 1 - e^{-\lambda x}$

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Моделирование процессов управления с учетом случайных факторов

1. Модель **биномиального распределения**.

ξ = число успехов в n независимых испытаниях с вероятностью успеха p

$$p_k = P(\xi = k) = C_n^k p^k (1 - p)^{n-k}$$

Для генерации случайной величины с биномиальным распределением имитируем n раз независимые испытания и подсчитываем число успехов:

```
k:=0;  
for i:= 1 to n do begin;  
x:=random;  
If x<p then k=k+1;
```

Легкий вариант: Excel ->
-> Анализ данных ->
-> Генерация случайных чисел

} Алгоритм модели
всё равно
надо знать

-Какой следующий этап моделирования?

Верификация модели, т.е. проверка, то ли распределение получилось ? А также оценить параметр p . Применяем методы математической статистики

Для этого необходимо сгенерировать много ($N=1000$) чисел (выборку) и посмотреть, как они распределены. Построить гистограмму на n ячеек и сравнить ее с идеальной гистограммой, т.е. для каждого k найти частоту попадания в k -ю ячейку $N p_k$

Как оценить вероятность p по выборке?

Среднее значение $M\xi = np$. По выборке x_1, x_2, \dots, x_N находим

выборочное среднее $\bar{x} = 1/N \sum_{i=1}^N x_i$, которое позволяет оценить p : $\hat{p} = \frac{\bar{x}}{nN}$

Как проверить распределение?

В зависимости от N есть два пути:

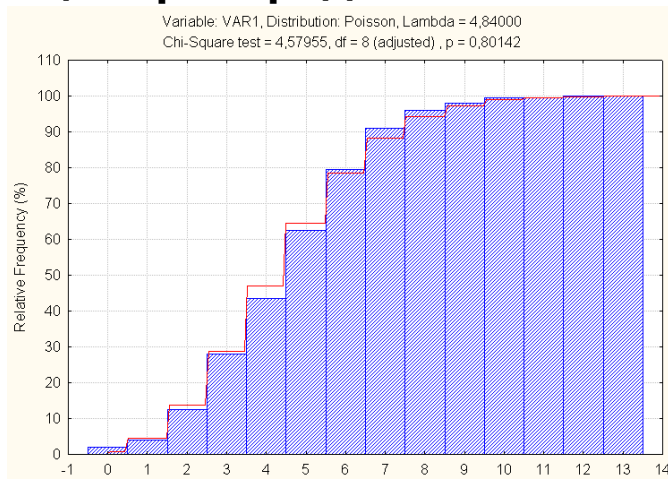
1. Для $N > 100$ получить **гистограмму**, т.е. таблицу, показывающую, какое число h_k ($k=1, 2, \dots, m$) элементов выборки попало в k -й интервал, и сравнить с тем, что должно быть теоретически по критерию χ^2

$$\chi^2 = \sum_{k=1}^m \frac{(h_k - np_k)^2}{np_k}$$

Здесь обозначено: n - общее число испытаний, m – число ячеек гистограммы, p_k - вероятность попасть в k -й интервал гистограммы

Эта статистика является **случайной величиной**, поскольку ее значение зависит от случайных значений h_k . Она распределена по закону χ^2 с m степенями свободы и для нее можно найти критическое значение $\chi^2_{крит}$, превышение которого происходит с малой вероятностью $\alpha=0.05$. Для $m > 20$ $\chi^2_{кр}$ можно вычислить по формуле $\chi^2_{кр} = m + 3\sqrt{2m}$. Если χ^2 , полученное по выборке, будет меньше $\chi^2_{кр}$, то моделирование прошло удачно.

2. Для $N \sim 50 \div 100$ следует упорядочить выборку по возрастанию $x(1) \leq x(2) \leq \dots \leq x(n)$. Такая упорядоченная выборка называется **вариационным рядом**. На основе вариационного ряда можно построить **эмпирическую функцию распределения**:



$$F_n(y) = \begin{cases} 0, & \text{если } y \leq X_{(1)}, \\ \frac{k}{n}, & \text{если } X_{(k)} < y \leq X_{(k+1)}, \\ 1 & \text{при } y > X_{(n)}. \end{cases}$$

$F_n(y)$ – кусочно-постоянная функция со скачками в точках $x(i)$. Если все $x(i)$ различны, то величина любого скачка равна $1/n$. $F_n(y)$ является случайной функцией, т.к. она зависит от случайных значений выборки.

Критерий Колмогорова основан на случайной величине $D_N = \sup |F_N(x) - F(x)|$. Критерием расхождения модельного и теоретического распределения служит Величина $\sqrt{ND_N}$, имеет асимптотическое распределение, найденное Колмогоровым, что позволяет найти критическое значение K_α . Для 95% -го уровня значимости ($\alpha=0.05$) $K_\alpha \approx 1.36$.

Действия в EXCEL

Генерация случайных чисел

Число переменных: 1

Число случайных чисел: 1000

Распределение: Биномиальное

Параметры

Значение p = 0,5

Число испытаний = 30

Случайное рассеивание:

Параметры вывода

Выходной интервал: \$A\$1:\$A\$1000

Новый рабочий лист:

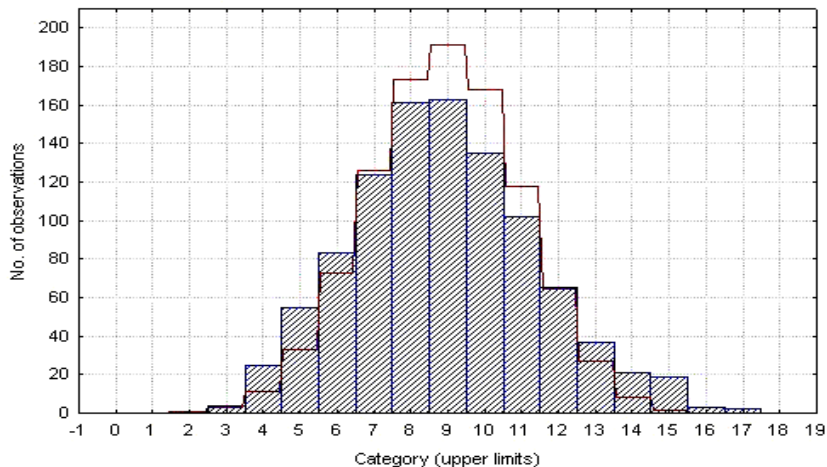
Новая рабочая книга

OK

Отмена

Справка

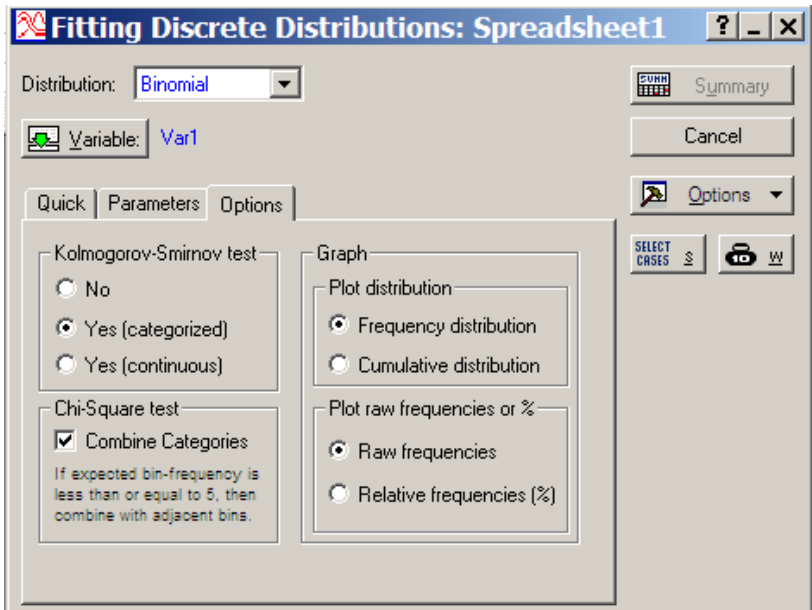
Variable: Var1, Distribution: Binomial, p = 0.52376
Chi-Square test = 155.94438, df = 9 (adjusted), p = 0.00000



Результат: 1000 случайных чисел в колонке А. Далее:

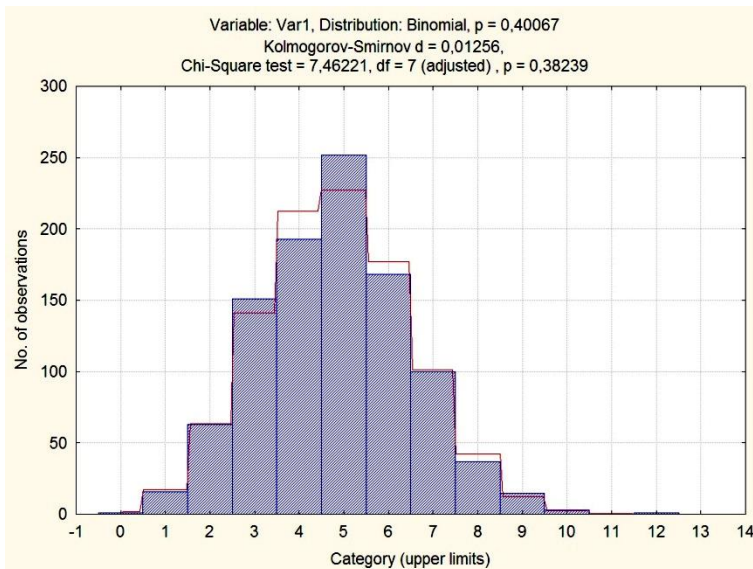
- вызвать программу **STATISTICA**
- Скопировать колонку А в столбец Var1 таблицы данных программы **STATISTICA**. Затем последовательно выполнить следующие действия:
 - выбрать в меню «Statistics->Distribution Fitting» («Статистика->Подгонка Распределений»);
 - на вкладке Discrete Distributions (Дискретные Распределения) выбрать Binomial (Биномиальное Распределение);
 - нажать ОК;
 - указать переменную Var1 после нажатия кнопки Variable;
 - нажать кнопку «Plot of observed and expected distribution» (График эмпирического и теоретического распределений).
 - перейти на вкладку Options (Опции) и в области кнопок Kolmogorov-Smirnov test выбрать пункт Yes (categorized);

Действия в STATISTICA



После того, как колонка А скопирована в столбец Var1 таблицы данных программы **STATISTICA**, последовательно выполнить следующие действия:

- выбрать в меню «Statistics->Distribution Fitting» («Статистика->Подгонка Распределений»);
- на вкладке Discrete Distributions (Дискретные Распределения) выбрать Binomial (Биномиальное Распределение);
- нажать ОК;
- указать переменную Var1 после нажатия кнопки Variable;
- перейти на вкладку Options (Опции) и в области кнопок Kolmogorov-Smirnov test выбрать пункт Yes (categorized);
- нажать кнопку «Quick», потом «Plot of observed and expected distribution» (График эмпирического и теоретического распределений).



Рекомендуемая литература

Основная

1.С.Г. Дмитриевский, Г.А.Ососков, Математическое моделирование часть 1, Учебно-методическое пособие, изд. Университет Дубна, 2011.

Электронный вариант есть на сайте <http://gososkov.ru/UNI-DUBNA/>
кликните для скачивания *handbook MathModelling.docm*

1.Алгазинов Э.К., Сирота А.А., Анализ и компьютерное моделирование информационных процессов и систем : Учебное пособие для вузов / М. : ДИАЛОГ-МИФИ, 2009,416с.

2.Ясницкий Л.Н., Введение в искусственный интеллект : Учебное пособие для студентов вузов / М. Академия, 2005, 176с.

3.Блаттер, К., Вейвлет-анализ. Основы теории : Учебное пособие для вузов / М. : Техносфера, серия: Мир математики, 2006. - 272с

Дополнительная

1. Гмурман В.Е., Теория вероятностей и математическая статистика, М.: Высш. шк. 2003, 479,

2. Емельянов А.А. Имитационное моделирование экономических процессов : Учебное пособие / 2-е изд., перераб. и доп. - : ИНФРА-М : Финансы и статистика, 2009. – 416 с.