# Статистика в системе компьютерной алгебры Mathematica 9

**Новая версия системы компьютерной алгебры Mathematica 9/9.01 фирмы Wolfram Research Inc. [1] существенно доработана и дополнена почти 400 новыми объектами (функциями). Значительное внимание при этом уделено вероятностным и статистическим вычислениям, реализующим аналитические (символьные) и численные методы с превосходными средствами графической визуализации. Поскольку возможности новой версии системы в нашей литературе вообще не рассматривались, настоящая статья восполняет этот пробел в области статистических вычислений и их полноцветной графической визуализации.**

**Владимир ДЬЯКОНОВ,**
**д. т. н., профессор**
vpdyak@yandex.ru

Статья предполагает, что читатель знаком с основами теории вероятностей и статистики [2], и лишь описывает мощные их инструментальные средства, присущие версии 9 системы Mathematica и частично трем ее предшествующим реализациям [3–6]. К достоинствам системы, прежде всего, следует отнести обширный центр документов с выходом в Интернет и с огромным числом демонстрационных примеров. На рис. 1 представлен в открытом виде раздел Probability&Statistics центра документов (справки) системы Mathematica 9. Вся информация в центре представлена на английском языке.

## Генерация случайных чисел

Наиболее часто статистические программы используются для генерации случайных чисел и массивов данных с ними. Для отдельных псевдослучайных реальных чисел предусмотрена функция ***RandomReal[]***. При каждом обращении к этой функции генерируется случайное (точнее, псевдослучайное) число в интервале от 0 до 1 с равномерным распределением. Функция ***RandomReal[xmin, xmax]*** генерирует случайное число в заданном интервале изменения переменной *x*. Есть и другие по синтаксису формы записи функции — ***RandomReal***. Функция используется как генератор псевдослучайных чисел и имитатор случайного шума. Примеры применения этой функции:

```
In[1]:= RandomReal[]
Out[1]= 0.157978
In[2]:= RandomReal[]
Out[2]= 0.399501
In[3]:= RandomReal[{−5, 5}]
Out[3]= 2.00357
In[4]:= RandomReal[{−5, 5}]
Out[4]= 4.6543
In[5]:= RandomReal[{−5, 5}]
Out[5]= 1.54934
In[6]:= RandomReal[{−5, 5}]
Out[6]= -4.10645
In[7]:= RandomReal[{−5, 5}, {3, 2}]
Out[7]= {{2.89308, −1.40616}, {−3.23635, −2.03801}, {−3.85406, −0.936425}}
```

Некоторые применения функции ***RandomReal*** представлены на рис. 2. Верхний рисунок является тестом на проверку равномерности распределения случайных чисел.

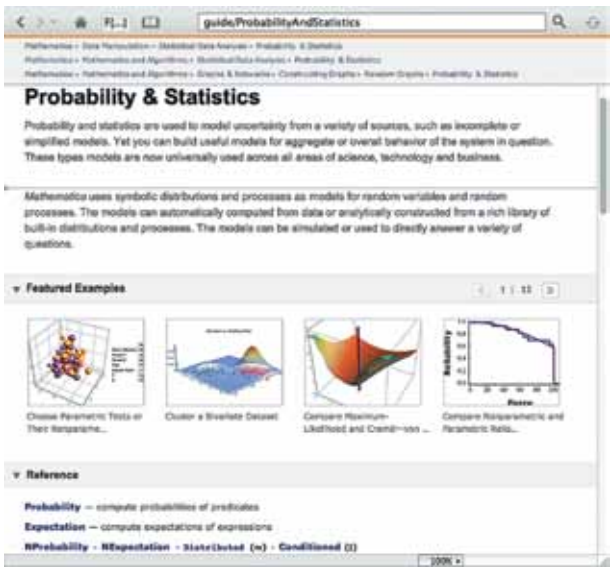


**Рис. 1.** Раздел Probability&Statistics центра документов системы Mathematica 9

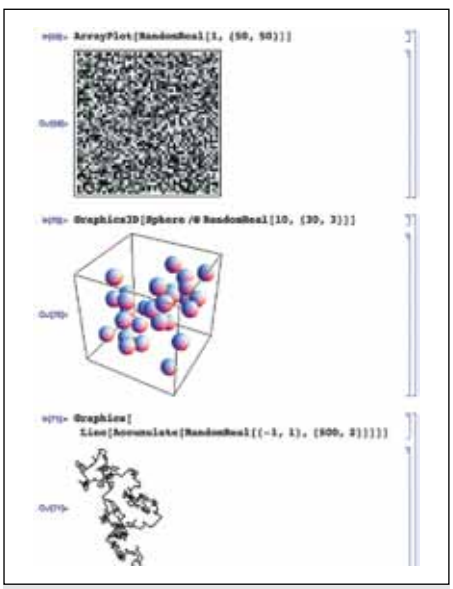


**Рис. 2.** Примеры применения функции RandomReal

Он демонстрирует случайный характер распределения чисел массива 50 на 50 чисел. Плотность расположения чисел практически одинакова по всему массиву. Второй рисунок показывает построение 30 шаров со случайными координатами их центров. И наконец, на третьем рисунке показана имитация броуновского движения частицы в пространстве.

Данные (непрерывные и дискретные) характеризуются рядом известных статистических параметров, таких как моменты, среднее (Mean), среднеквадратическое отклонение (StandardDerivation), медиана (Median), квантиль (Quantile), гармоническое среднее (HarmonicMean) и др. Отличительной чертой Mathematica 9 является возможность вычисления их в аналитическом (символьном) виде:

```
In[1]:= Mean[{a, b, c, d, e}]
Out[1]= 1/5 (a + b + c + d + e)
In[2]:= Mean[{{a, u}, {b, v}, {c, w}}]
Out[2]= {1/3 (a + b + c), 1/3 (u + v + w)}
In[3]:= Mean[LogNormalDistribution[μ, 1]]
Out[3]= E^(1/2 + μ)
In[4]:= Variance[LogNormalDistribution[0, 1]]
Out[4]= (−1 + E) E
In[5]:= StandardDeviation[LogNormalDistribution[0, 1]]
Out[5]= Sqrt[(−1 + E) E]
In[6]:= Median[ExponentialDistribution[λ]]
Out[6]= Log[2]/λ
In[7]:= Quantile[NormalDistribution[μ, σ], q]
Out[7]= μ – Sqrt[2] σ InverseErfc[2 q]
In[8]:= HarmonicMean[1/Log[{a, b, c, d}]]
Out[8]= 4/(Log[a] + Log[b] + Log[c] + Log[d])
```

Разумеется, возможно и вычисление численных данных:

```
In[1]:= Mean[{1.21, 3.4, 2.15, 4, 1.55}]
Out[1]= 2.462
In[2]:= GeometricMean[{{5, 10}, {2, 1}, {4, 3}, {12, 15}}]
Out[3]= {2 30^(1/4), 2^(1/4) Sqrt[15]}
In[4]:= data = RandomVariate[NormalDistribution[], 10^3];
In[4]:= D = SmoothKernelDistribution[data];
In[5]:= Moment[D, 2]
Out[5]= 1.04922
In[6]:= Quantile[F, 0.95]
Out[6]= 1.70912
```

Наиболее часто встречается нормальное распределение вероятностей. Графики функций нормального распределения вероятностей от одной и двух переменных показаны на рис. 3. Для представления функции двух переменных используется трехмерный. Графики могут изменяться в масштабах, сохраняя неизменным соотношение сторон. А трехмерные графики можно поворачивать с помощью мыши.

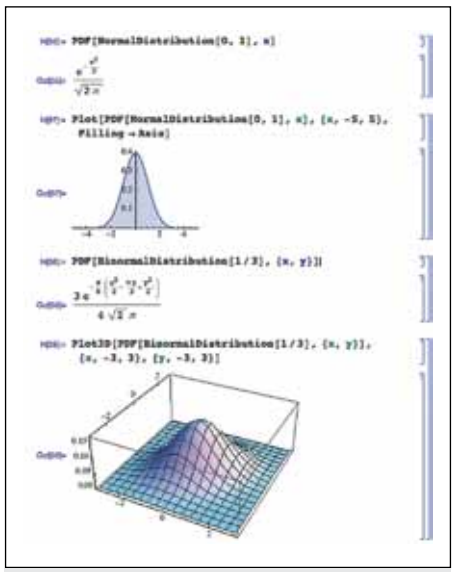

**Рис. 3.** Функции нормального распределения вероятностей от одной и двух переменных

Графики функций распределения вероятностей часто используются для представления результатов статистических тестов. Mathematica 9 позволяет задать любую функцию распределения вероятностей и имеет множество встроенных функций для наиболее типовых распределений: экспоненциального, γ, *F*, $\chi^2$, биномиального, Пирсона и др.

Функции характеризуются различными моментами и другими параметрами. Не будем на них останавливаться детально, так как это общеизвестные параметры и Mathematica 9 позволяет получить для них формулы в аналитическом виде, а при необходимости и построить график того или иного параметра (рис. 4).

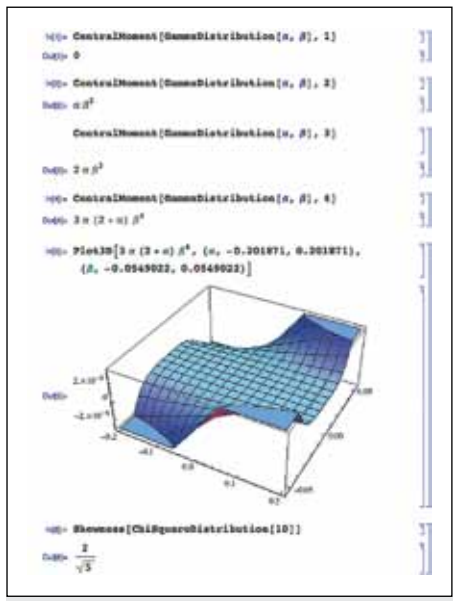

**Рис. 4.** Центральные моменты для γ-функции распределения и асимметрия для $\chi^2$-функции

Количество функций распределения вероятностей для непрерывных и дискретных данных в Mathematica 9 очень велико, и они охватывают большинство известных функций вероятностей. Некоторые принадлежат к определенному классу функций, что отмечено в их теоретическом описании в справке. Для таких функций приводятся диаграммы соответствия, например диаграмма, показанная на рис. 5.

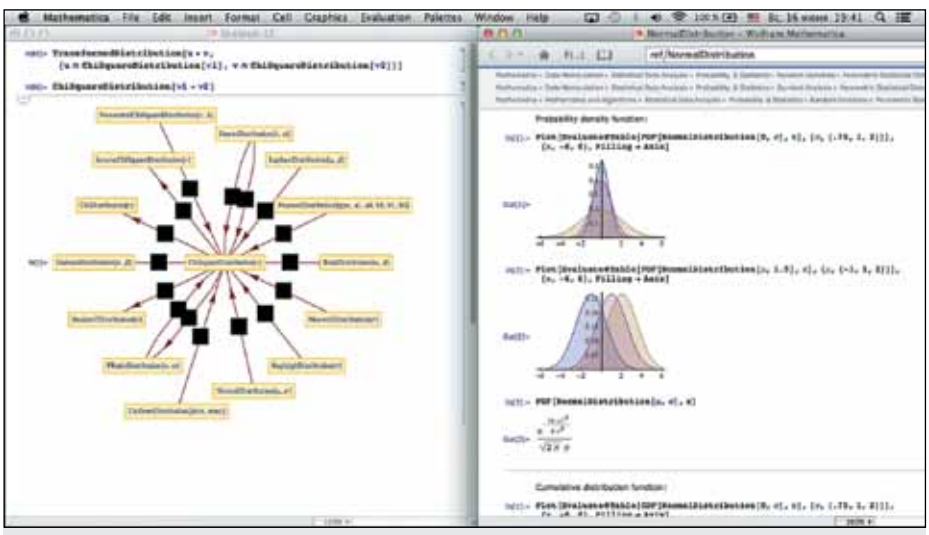

**Рис. 5.** Функции распределения вероятности группы $\chi^2$

## Основные виды статистической графики

Статистические данные обычно представляются графиками специального вида. Теория вероятности и статистика конца 1980-х годов не были избалованы графическими иллюстрациями. Например, на сотни страниц текста и формул в [2] можно отыскать лишь с десяток простых графиков.

Но Mathematica 9 имеет обширные возможности в построении цветных графиков, обычно четких и аккуратных, отличающихся высоким полиграфическим качеством. Окно справки с указанием основных функций статистической графики показано на рис. 6 и позволяет быстро подобрать нужный тип графика. Много графиков специального назначения, применимых и в статистике, есть и в разделах справки, посвященных финансовым расчетам.

Массивы данных часто представляются гистограммами по столбцам или строками. Если массив одномерный — его представляют столбиковыми гистограммами. При этом данные по горизонтали разбиваются на N участков по числу столбцов (или строк) будущей гистограммы. В каждом из частичных участков подсчитывается сумма данных — число, и оно определяет высоту (или длину) столбца. Таким образом, проводится первичная статистиче-


**Рис. 6.** Окно с основными функциями статистической графики

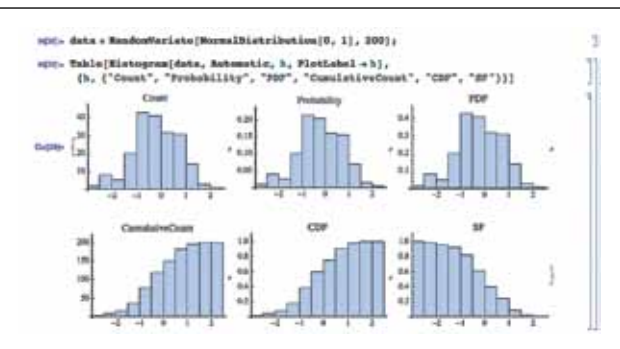


**Рис. 7.** Столбиковые гистограммы статистических параметров

ская обработка данных. На рис. 7 показано построение ряда простых столбиковых гистограмм. Возможна их закраска разными цветами.

Столбиковые гистограммы возможны различного типа: двойные и множественные, построенные кружками, а не прямоугольниками, окрашенные в разные оттенки цвета и т. д. (рис. 8). Как правило, тип таких гистограмм имеет чисто косметическое значение, но может и указывать на тот или иной вид объектов.

Широко применяются и круговые гистограммы различного типа — от разбитого на секторы круга или диска до представления отдельных секторов (рис. 9). Полный круг обычно соответствует сумме всех значений чисел, которая принимается за 100%. От нее рассчитываются размеры секторов. Круговые гистограммы могут быть плоскими и объемными (3D-типа).

## Сглаживание данных и очистка от шума

Статистические методы часто используются для сглаживания данных и очистки сигналов от шума. Наиболее часто применяются методы усреднения с плавающим окном. Mathematica позволяет получить аналитические выражения для формул сглаживания:

```
In[1]:= MovingAverage[{a, b, c, d, e}, 2]
Out[1]= {(a + b)/2, (b + c)/2, (c + d)/2, (d + e)/2}
In[2]:= MovingMedian[{1, E, Sqrt[Pi], 2, 10, E^2}, 3]
Out[2]= {Sqrt[π], 2, 2, E^2}
In[3]:= ExponentialMovingAverage[{a, b, c}, x]
Out[3]= {a, a + (–a + b) x, a + (–a + b) x + x (–a + c – (–a + b) x)}
In[4]:= MeanFilter[ {a, b, c}, 1]
Out[4]= {(a + b)/2, 1/3 (a + b + c), (b + c)/2}
```

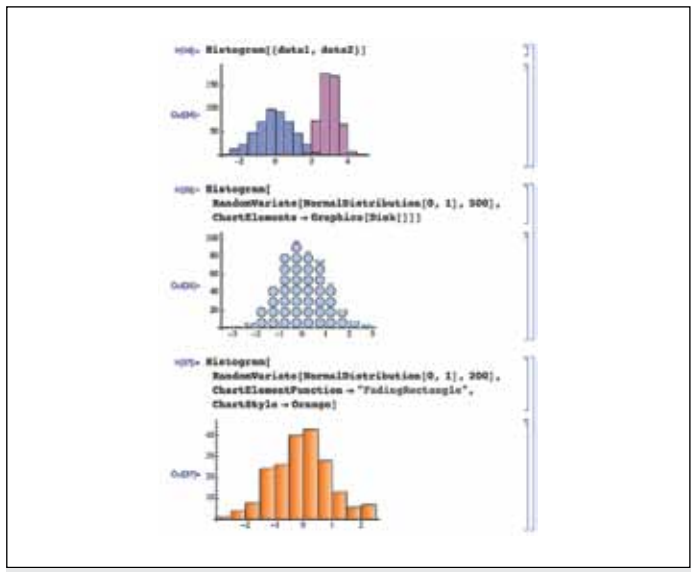


**Рис. 8.** Специальные типы столбиковых гистограмм

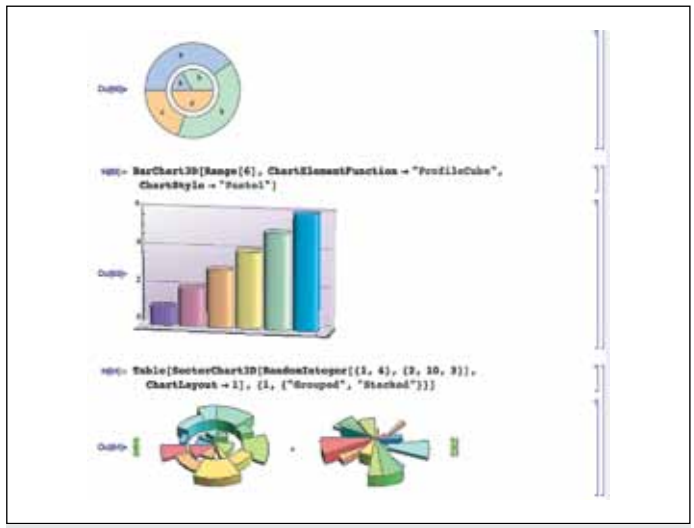


**Рис. 9.** Круговые гистограммы

Наибольшую степень сглаживания обычно обеспечивает экспоненциальное сглаживание.

## Корреляция и свертка

Корреляция и свертка — еще два хорошо известных понятия статистики. Обычно их относят к двум или нескольким наборам данных (массивов). Часто рассматривается корреляция и свертка некоторого ядра и данных или двух векторов или матриц:

```
In[1]:= Correlation[{a, b}, {x, y}]
Out[1]= ((a – b) (Conjugate[x] – Conjugate[y]))/(Sqrt[(a – b) (Conjugate[a] – Conjugate[b])]
Sqrt[(x – y) (Conjugate[x] – Conjugate[y])])
In[2]:= Correlation[{1.5, 3, 5, 10}, {2, 1.25, 15, 8}]
Out[2]= 0.475976
In[3]:= ListConvolve[{x, y}, {a, b, c, d, e}]
Out[3]= {b x + a y, c x + b y, d x + c y, e x + d y}
In[4]:= ListCorrelate[{x, y}, {a, b, c, d, e}]
Out[4]= {a x + b y, b x + c y, c x + d y, d x + e y}
```

Корреляция характеризует степень идентичности их изменения (рис. 10, сверху). Если коэффициент корреляции равен 0, то взаимосвязь наборов данных отсутствует, а если он равен 1, то данные меняются одинаковым образом (например, если они представляют парал-

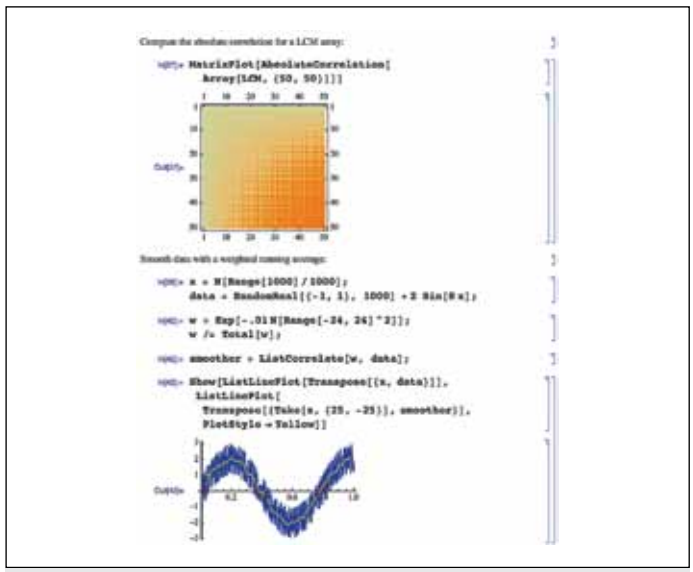

**Рис. 10.** Абсолютная корреляция для LCM-массива и пример сглаживания зашумленного сигнала и его очистки от шума

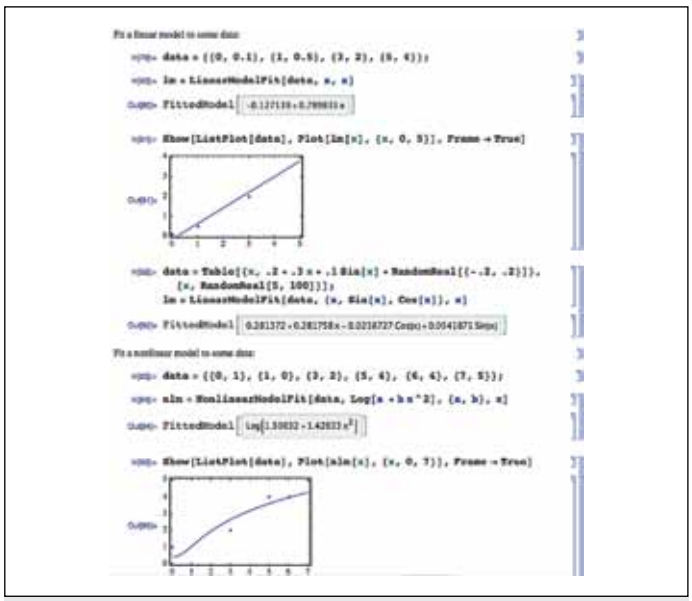

**Рис. 11.** Линейная и нелинейная модели регрессии для одномерного массива данных

лельные прямые). Корреляция и свертка широко применяются для сглаживания сигналов и очистки их от шума (рис. 10, внизу).

## Линейная и нелинейная регрессия

Из статистических методов обработки данных особенно часто применяется регрессия, основанная на методе наименьших квадратов. Mathematica 9 имеет прекрасный и довольно универсальный аппарат для проведения линейной и нелинейной регрессии, реализованный всего несколькими достаточно универсальными функциями. Простейшей из них является функция:

```
Fit[data, funs, vars]
```

которая для данных ***data*** реализует метод наименьших квадратов как линейную комбинацию функций ***funs*** переменных ***vars*** (линейная регрессия общего вида). Примеры:

```
In[50]:= data = {{0, 1}, {1, 0}, {3, 2}, {5, 4}};
In[51]:= line = Fit[data, {1, x}, x]
Out[51]= 0.186441 + 0.694915 x
In[52]:= parabola = Fit[data, {1, x, x^2}, x]
Out[52]= 0.678392 – 0.266332 x + 0.190955 x^2
In[53]:= data = {{–Pi, 4}, {–Pi/2, 0}, {0, 1}, {Pi/2, –1}, {Pi, -4}};
In[54]:= Fit[data, {Sin[x/2], Sin[x], Sin[2 x]}, x]
Out[54]= 0. – 4. Sin[x/2] + 2.32843 Sin[x]
In[55]:= data1 = {{0, 0, 0},{1, 0, 1},{0, 1, 2},{1, 1, 0},{1/2, 1/2, 1}};
In[56]:= plane = Fit[data1, {1, x, y}, {x, y}]
Out[56]= 0.8 – 0.5 x + 0.5 y
```

Сами функции могут быть и нелинейными, и от нескольких переменных. Но их комбинация должна быть линейной. На рис. 11 показаны линейная и нелинейная модели регрессии для одномерного массива данных с выводом их графиков и исходных точек.

Синусоидальная модель регрессии (рис. 12) обеспечивает построение синусоиды, проходящей в облаке исходных точек с наименьшим отклонением от каждой из них по методу наименьших квадратов. Следует отметить, что при малом числе точек по виду их облака трудно или даже невозможно судить о принадлежности точек и лишь проведение регрессии позволяет установить, к какой функции принадлежат точки.

Иногда данные желательно представлять достаточно простой функцией определенного класса. В таких случаях обычно применяется полиномиальная модель регрессии. На рис. 13 показана программа полиномиальной регрессии, использующая средства динамической

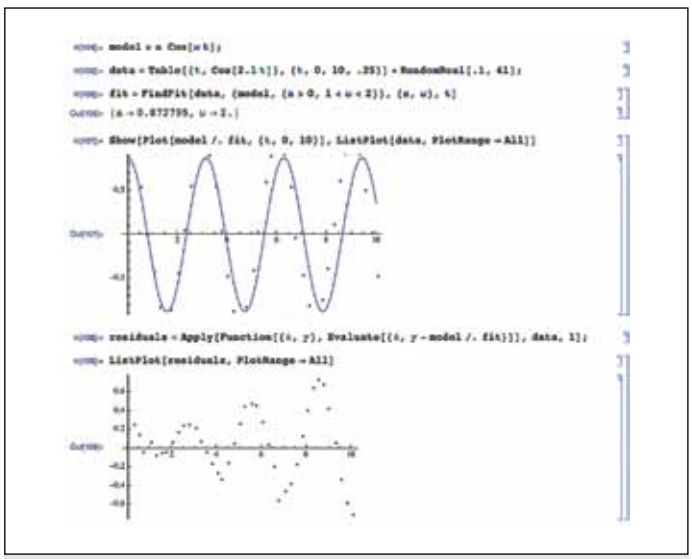

**Рис. 12.** Синусоидальная модель нелинейной регрессии для одномерного массива данных

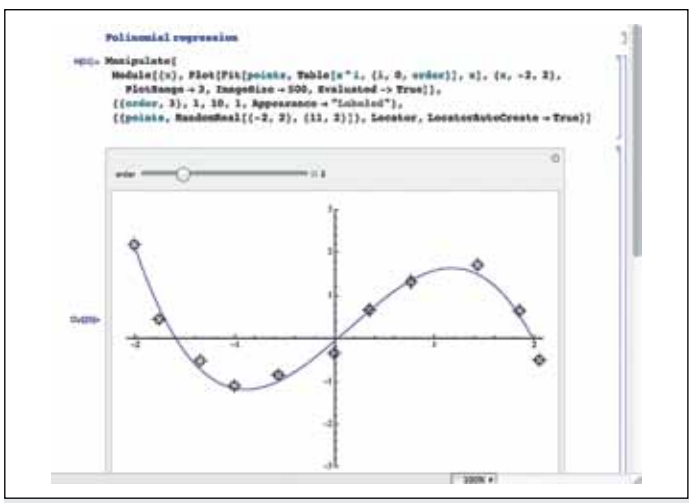

**Рис. 13.** Полиномиальная регрессия и аппроксимация с динамической интерактивностью

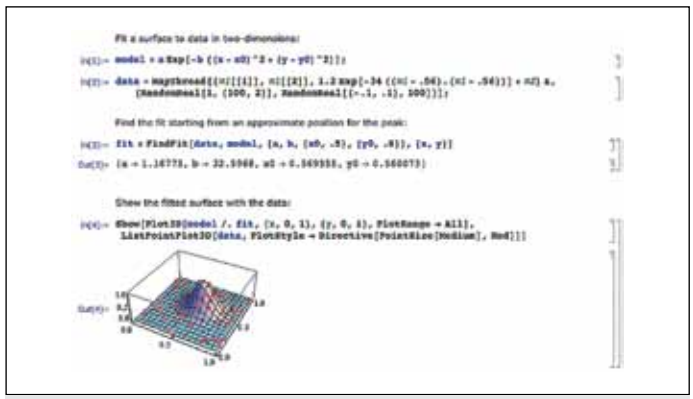


**Рис. 14.** Экспоненциальная нелинейная модель регрессии для функции двух переменных

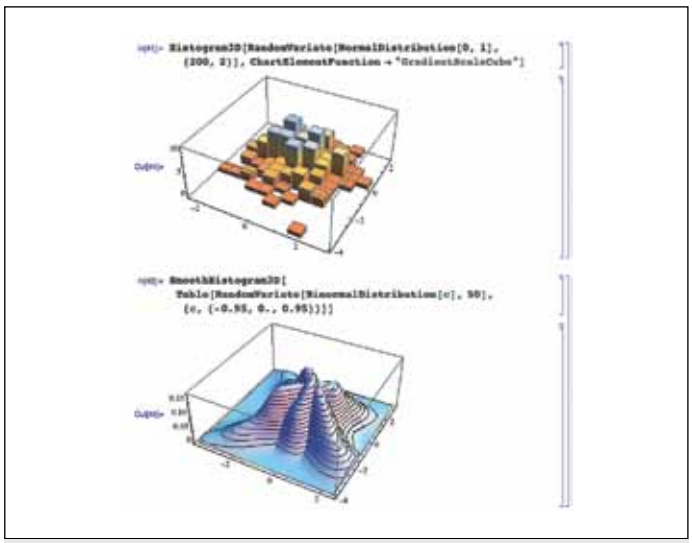

**Рис. 18.** Дискретная и сглаженная трехмерные гистограммы

интерактивности. Исходные точки данных располагаются на плоскости, и их можно произвольно перемещать мышью с нажатой левой клавишей. Кроме того, слайдером можно менять степень полинома *n*.

Меняя набор данных, можно следить за изменением функции регрессии (аппроксимации). Если степень полинома на 1 меньше числа точек, то программа дает полиномиальную аппроксимацию, при которой график функции аппроксимации точно проходит через все заданные точки. В данном случае число точек равно 11, но их число можно изменить как параметр функции ***RandomReal*** в программе.

Модели регрессии могут работать и с двумерными и многомерными данными. На рис. 14 представлена модель экспоненциальной регрессии для данных, приближаемых функцией регрессии, с двумя переменными. На графике представлены в пространстве исходные случайные точки и приближающая их поверхность. Для ее построения используется 3D-графика.

## Представление специальных данных

В Mathematica 9 могут обрабатываться самые разнообразные специальные данные о странах и городах, финансах, физических явлениях и т. д. Географические данные о странах мира содержат стилизованные карты страны, данные о численности населения, положение на карте мира (атласе) и т. д. Данные о Российской Федерации представлены на рис. 15. Следует отметить, что сведения по ряду стран в базе данных самой системы отсутствуют. Они хранятся на интернет-сайте фирмы Wolfram Research Inc. и по запросу считываются с сайта. Это может занять несколько секунд и сопровождаться комментариями. Аналогичным образом можно получить и данные по крупным городам разных стран. Примеры представления данных о некоторых городах показаны на рис. 16.

Довольно часто приходится работать с финансовыми данными — о доходах и расходах населения, курсах акций и валют и т. д. Одна из функций для удобного представления подобных данных — DateListPlot. Примеры ее применения представлены на рис. 17.

Для двумерных данных часто используются трехмерные гистограммы. Они могут быть дискретными или сглаженными. Пример построения таких гистограмм показан на рис. 18.

Иногда для таких данных используются гистограммы плотности, в том числе с контурными линиями и цветовой окраской (рис. 19). Большие возможности в представлении двумерных данных дает функция ListPlot3D. Наряду с формируемыми математически изображениями тут можно составлять комбинированные изображения — например, лицо (рис. 20).

## Представление реальных объектов

Часто возникает задача построения графиков, достаточно хорошо воспроизводящих вид реальных объектов. На рис. 21, к примеру, дано построение графика молекулы протеина и представлена таблица

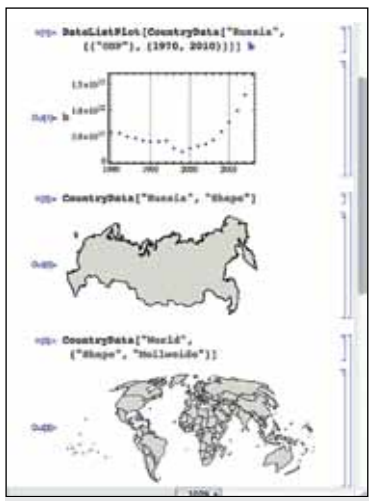

**Рис. 15.** Географические данные о странах мира

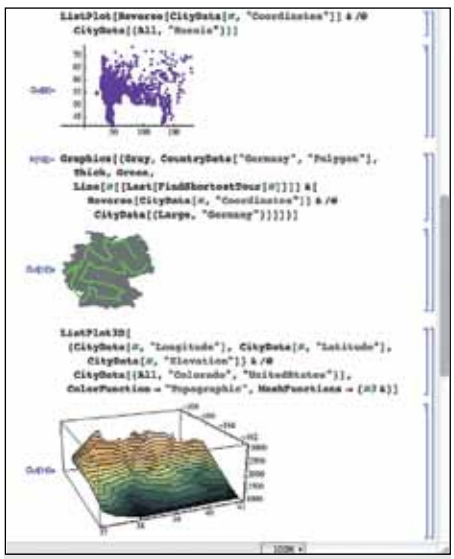

**Рис. 16.** Географические данные о городах

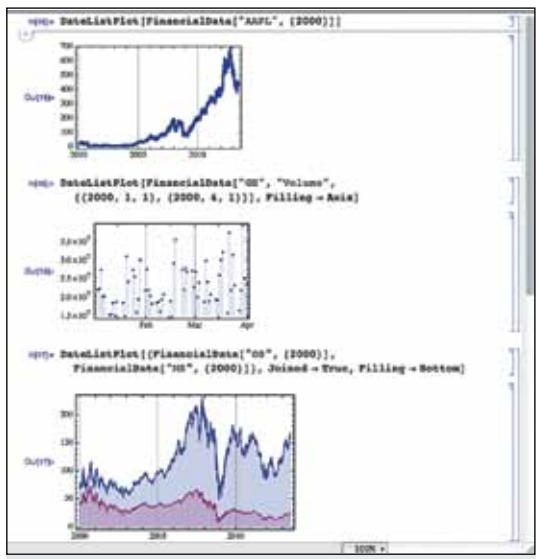

**Рис. 17.** Примеры представления финансовых данных

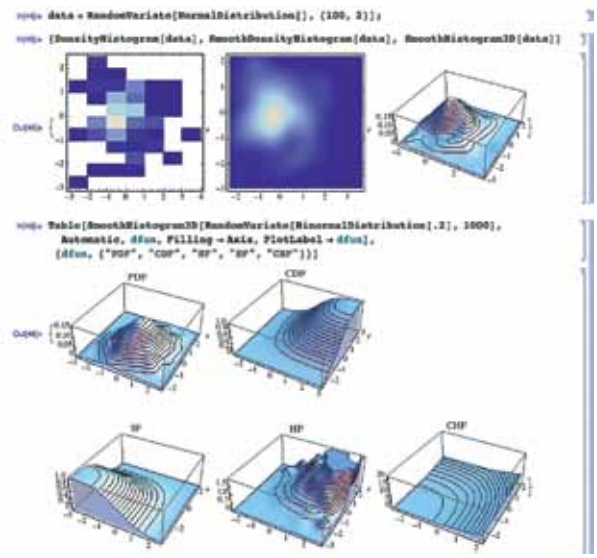


**Рис. 19.** Трехмерные гистограммы плотности и с контурными линиями

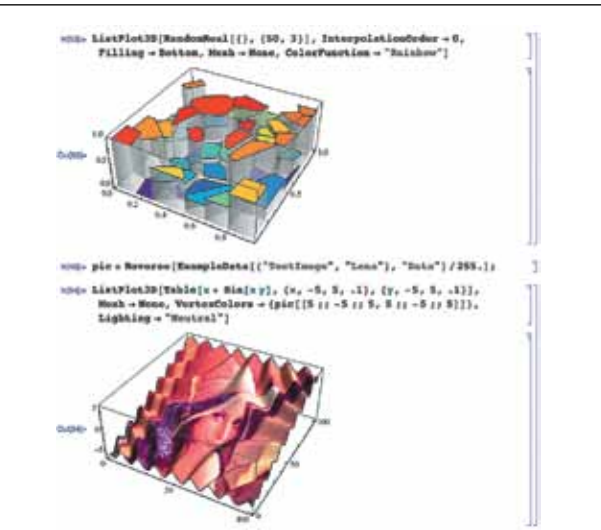

**Рис. 20.** Примеры применения функции ListPlot3D с окраской

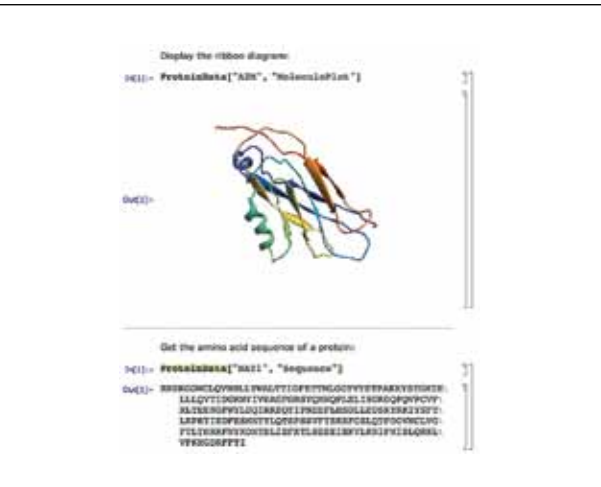

**Рис. 21.** Молекула протеина и последовательность ее генных кодов

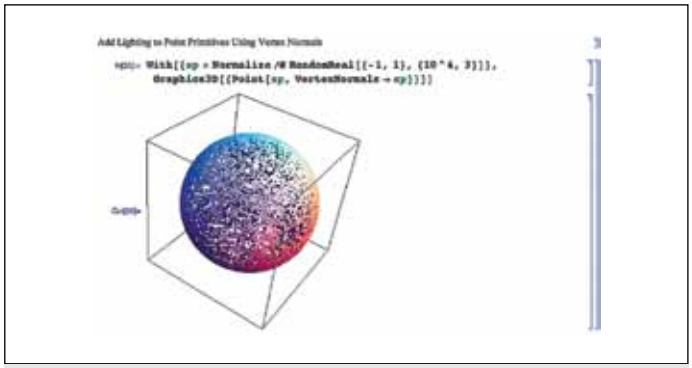

**Рис. 22.** Шар с окраской Vertex Normal

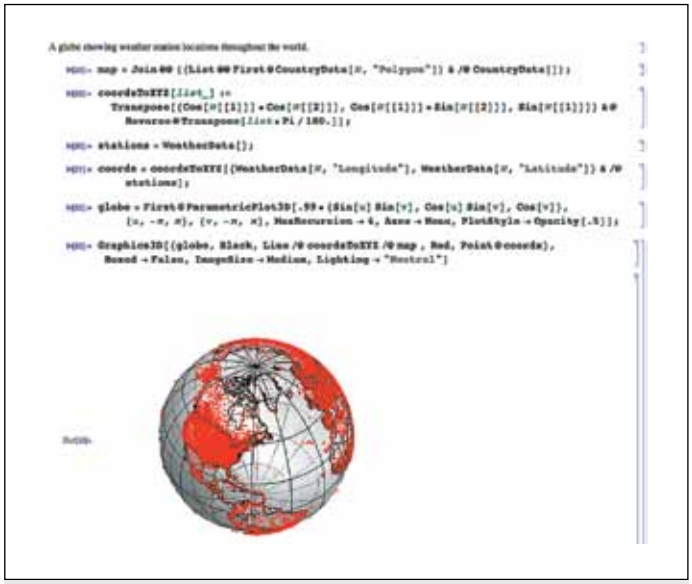

**Рис. 23.** Земной шар со станциями погоды

ее генных кодов. Подобные данные широко применяются в генной микробиологии.

Шар с характерной окраской Vertex Normal строит программа, показанная на рис. 22. Окраска придает шару шероховатый вид.

А на рис. 23 показана программа, выводящая изображения земного шара (глобуса) с нанесенными на его поверхность станциями погоды. Нетрудно заметить, что они расположены очень неравномерно.

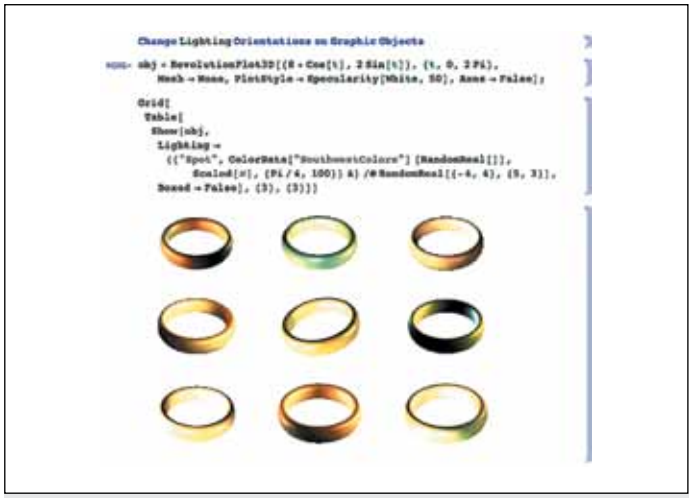

**Рис. 24.** Изображения колец со случайной окраской при освещении

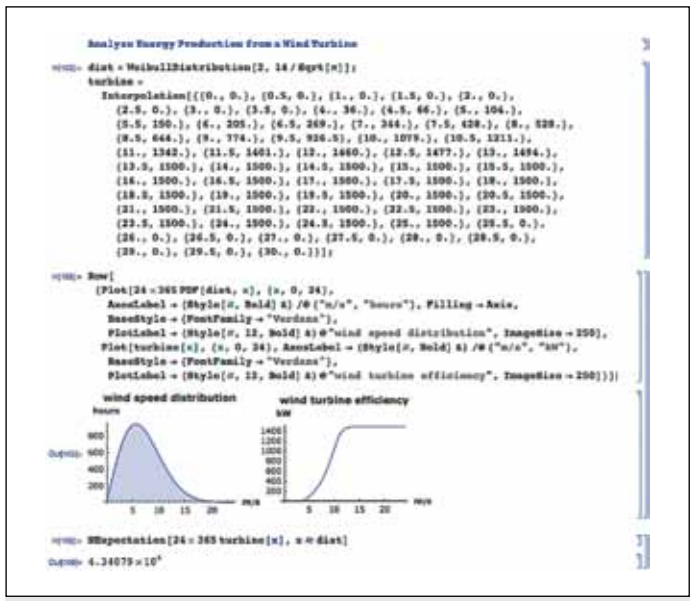


**Рис. 25.** Представление данных о скорости и мощности турбины ветровой электростанции

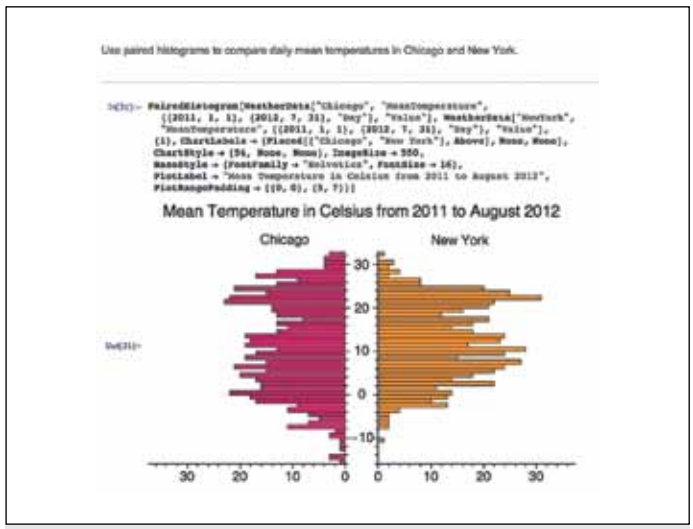


**Рис. 26.** Сравнение данных о средней температуре в двух городах США

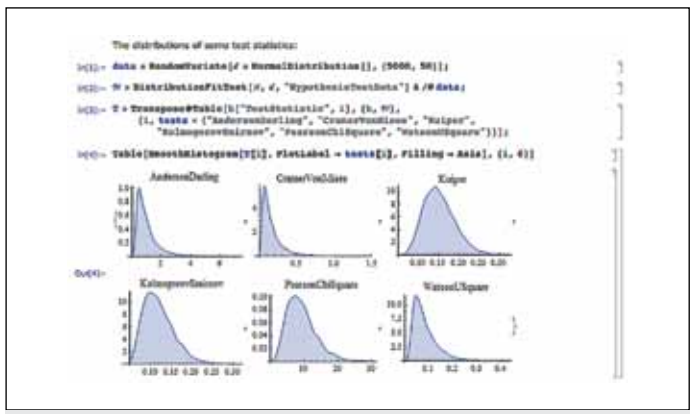


**Рис. 27.** Функции распределения для некоторых статистических тестов

На рис. 24 показаны изображения колец, имитирующие различную случайную окраску при освещении от внешнего источника света. Кольца выглядят очень реалистично.

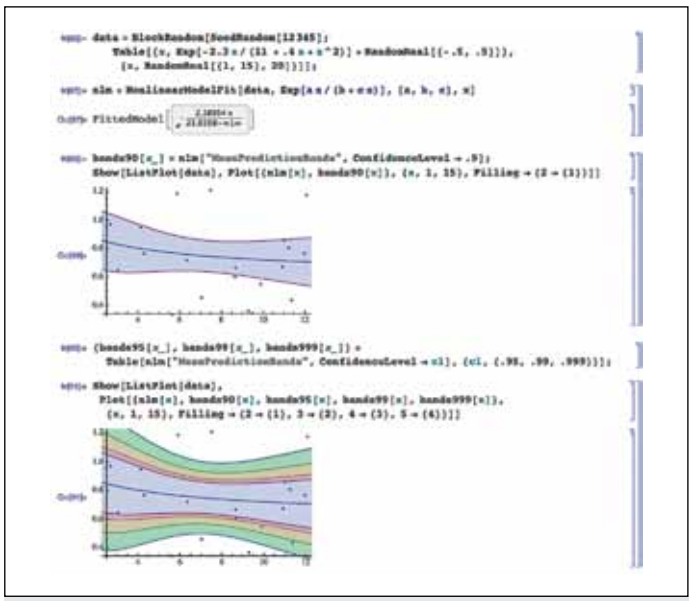

**Рис. 28.** Экспоненциальная регрессия с доверительными интервалами

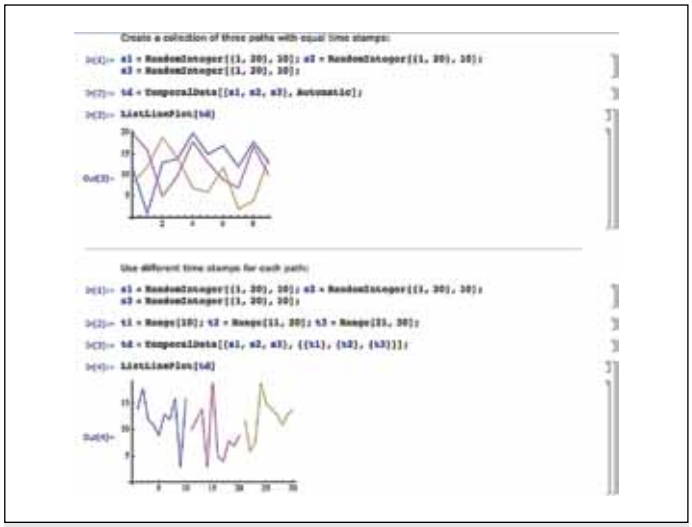


**Рис. 29.** Представление временных рядов

## Решение реальных статистических задач

На рис. 25 показано представление данных о скорости вращения и мощности турбины одной из ветровых электростанций. Исходные данные представлены в массиве, и программа обеспечивает их выделение и вывод с интерполяцией и сглаживанием.

Другой пример — сравнение данных о средней температуре в двух городах — Чикаго и Нью-Йорке. В этом случае удобнее оказались сдвоенные горизонтальные гистограммы. Левая гистограмма на рис. 26 относится к одному городу, правая — к другому.

Многие статистические тесты основаны на анализе сглаженных функций распределения и их параметров. Часто для этого достаточно построить сглаженную функцию распределения того или иного теста (рис. 27).

Экспоненциальная регрессия с доверительными интервалами показана на рис. 28. Доверительные интервалы указывают области, в которых данные имеют достаточную надежность.

Многие статистические данные удобно представлять в виде временных рядов. Работа с ними впервые включена в ядро системы Mathematica 9. На рис. 29 показано построение целочисленных временных рядов с помощью функции ***TemporalData*** на трех разных временных интервалах.

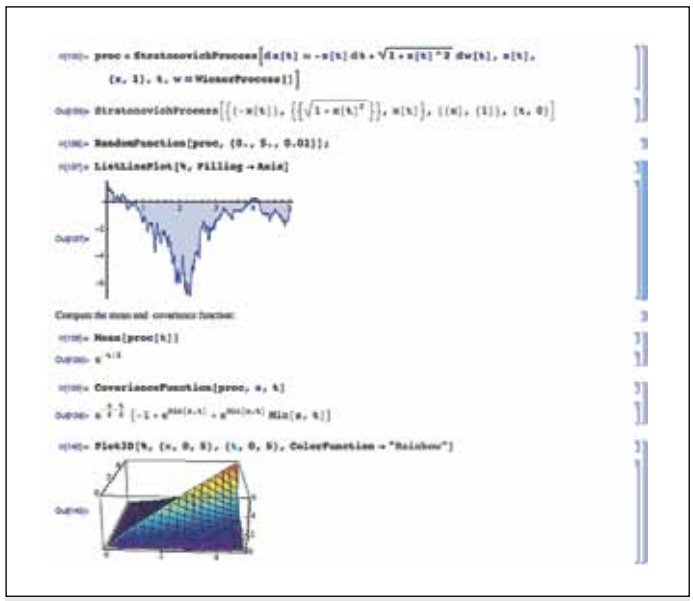

**Рис. 30.** Представление стохастического процесса и ковариационной функции

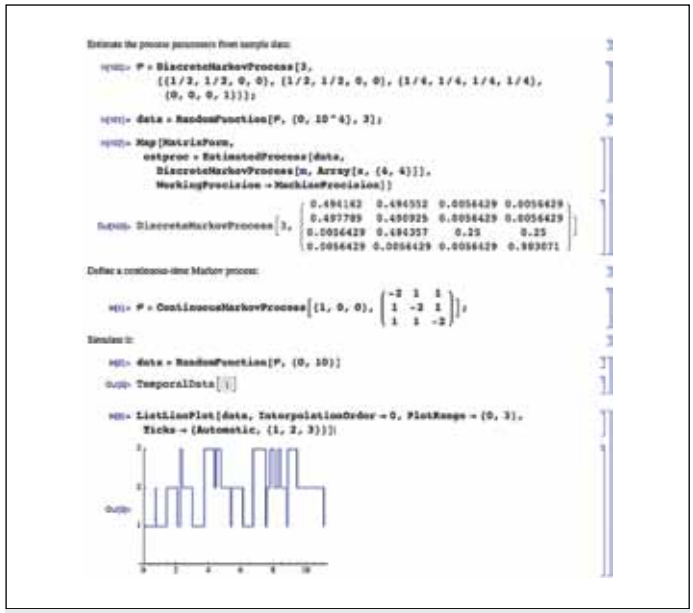

**Рис. 31.** Дискретный и непрерывный Марковские процессы

В теории и практике статистики есть множество стохастических (случайных) процессов с различными функциями распределения и различными их параметрами. Задание и представление случайного процесса Стратоновича показано на рис. 30 (сверху). В нижней части рис. 30 представлено вычисление среднего и ковариационной функции этого процесса и построение трехмерного графика этой функции. В справке можно найти множество примеров, иллюстрирующих и иные виды случайных процессов, например случайные процессы Винера.

Интересными свойствами обладают Марковские процессы, при известном «настоящем» их «прошлое» и «будущее» не зависят друг от друга. Примером может служить распад радиоактивного элемента. Такие процессы могут быть дискретными и непрерывными (рис. 31).

## Кластеризация

Данные часто группируются в некоторых областях, объединяющих их по определенным признакам. Такие области данных называют кластерами (в пер. claster — «гроздь»).

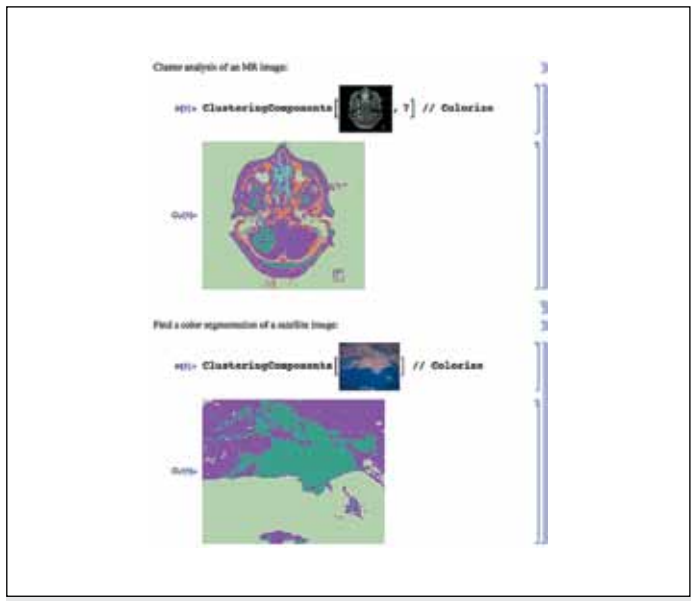

**Рис. 32.** Кластеризация изображения среза черепной коробки и данных со спутника

На рис. 32 сверху показана кластеризация для изображения среза черепной коробки человека. Выявление кластеров в данном случае облегчает анализ среза и иногда выявляет опухоли и иные аномальные образования. Хорошие результаты кластеризация дает также при анализе данных, поступающих от спутников Земли, и от данных аэрофотосъемки (рис. 32, внизу).

## Моделирование ARMA-процессов

Одной из математических моделей, использующихся для анализа и прогнозирования стационарных временных рядов в статистике, является модель ARMA. Она обобщает две более простые модели временных рядов — модель авторегрессии (AR) и модель скользящего среднего (MA). Пример моделирования ARMA-процесса представлен на рис. 33.

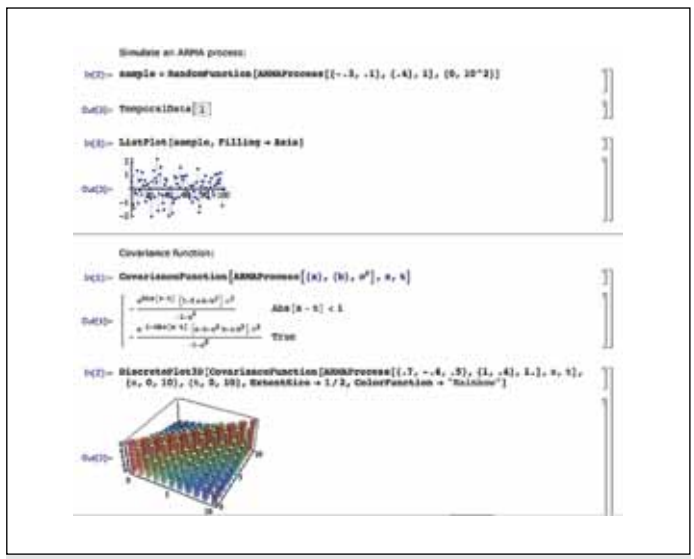

**Рис. 33.** Моделирование ARMA-процесса

## Дисперсионный анализ

Дисперсионный анализ в Mathematica 9 осуществляет специальный пакет ANOVA. В простейшем случае дисперсионный анализ данных

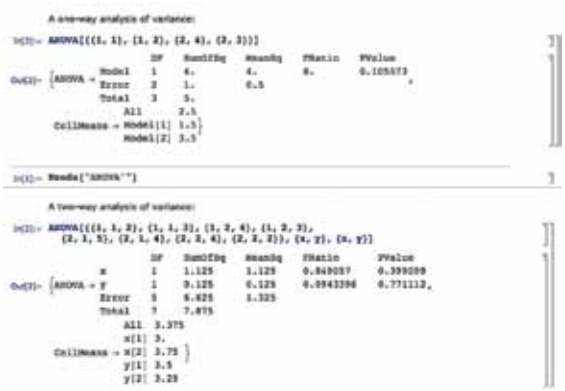


**Рис. 34.** Примеры дисперсионного анализа

data осуществляет функция ANOVE[data] (рис. 34). Результаты дисперсионного анализа выдаются в табличной форме.

## Экспоненциальная многокомпонентная регрессия с анализом компонент

Иногда при проведении регрессии приходится выбирать функцию, содержащую несколько компонент, и анализировать или сравнивать параметры этих компонент. Такой пример экспоненциальной регрессии представлен на рис. 35.

## Заключение

В Mathematica 9 осуществлен прорыв в области статистических вычислений и их графической визуализации в виде красочных цветных рисунков типографического качества. Система лидирует по числу статистических функций, встроенных в ядро, и в отличие от электронных таблиц и статистических программ обеспечивает смешанную аналитическую и числовую парадигму вычислений.

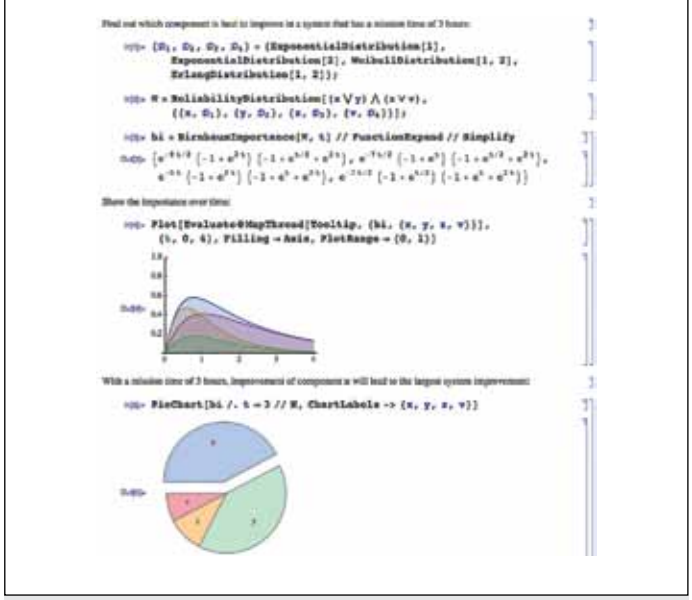


**Рис. 35.** Экспоненциальная многокомпонентная регрессия

## Литература

1. www.wolfram.com
2. Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф. Справочник по теории вероятностей и математической статистике. М.: Наука. Физматлит. 1985.
3. Дьяконов В. П. Mathematica 5.1/5.2/6 в научно-технических расчетах. Издание 2-е переработанное и дополненное. М.: Солон-Пресс. 2009.
4. Дьяконов В. П. Mathematica 5/6/7. Полное руководство. С.: ДМК-Пресс. 2009.
5. Дьяконов В. П. Задание, анализ и обработка сигналов в системе Mathematica 8 // Компоненты и технологии. 2012. № 8.
6. Дьяконов В. П. Обработка изображений в СКМ Mathematica 8 // Компоненты и технологии. 2012. № 10.