Тема №8 «Математическая статистика. Обработка данных»

Лекция 1 «Основы программирования случайных величин, характеристик и процессов»

1. Моделирование законов распределения
2. Статистические характеристики. Построение гистограмм.
3. Случайные процессы

---

1. Моделирование законов распределения

Для моделирования различных физических, и прочих эффектов широко распространены методы, называемые методами Монте-Карло. Их основная идея состоит в создании определенной последовательности случайных чисел, моделирующей тот или иной эффект, например, шум в физическом эксперименте, случайную динамику биржевых индексов и т. п. Для этих целей в Mathcad имеется ряд встроенных функций, реализующих различные типы генераторов псевдослучайных чисел.

Mathcad имеет большой набор функций математической статистики, позволяющих вычислять характеристики выборки данных: средние величины, дисперсию, коэффициенты корреляции и другие коэффициенты, вычислять плотности вероятности, функции вероятности, квантили вероятности для 17 различных видов распределения случайных величин. Кроме того, в нем есть генераторы случайных чисел, соответствующие всем 17 видам распределения.

Согласно определению, случайная величина принимает то или иное значение, но какое конкретно, зависит от случайных обстоятельств опыта и заранее точно предсказано быть не может. Можно лишь говорить о вероятности $P(X)$ принятия случайной дискретной величиной того или иного значения $x_к$, или о вероятности попадания непрерывной случайной величины в тот или иной числовой интервал $(x,x+dx)$. Вероятность $P(X)$, может принимать значения от 0 до *1* (случайная величина заведомо примет значение от x до $x+dx$). Соотношение $F(X)$ называют законом распределения случайной величины, а зависимость $W(X)$ между возможными значениями непрерывной случайной величины и вероятностями попадания в их окрестность называется ее плотностью вероятности (**probability density**).

В Mathcad имеется ряд встроенных функций, задающих используемые в математической статистике законы распределения. Они вычисляют как значение плотности вероятности различных распределений по значению случайной величины x, так и некоторые сопутствующие функции. Все они, по сути, являются либо встроенными аналитическими зависимостями, либо специальными функциями. Большой интерес представляет наличие генераторов случайных чисел, создающих выборку псевдослучайных данных с соответствующим законом распределения.

Рассмотрим подробно возможности Mathcad на нескольких наиболее популярных законах распределения, а затем приведем перечень всех распределений, встроенных в Mathcad.

Нормальное (Гауссово) распределение

В теории вероятности доказано, что сумма различных *x* зависимых случайных слагаемых (независимо от их закона распределения) оказывается случайной величиной, распределенной согласно нормальному закону (т. н. центральная предельная теорема). Поэтому нормальное распределение хорошо моделирует самый широкий круг явлений, для которых известно, что на них влияют несколько независимых случайных факторов.

Перечислим основные встроенные функции, имеющиеся в Mathcad для описания нормального распределения вероятностей:

- **dnorm(x,m,o)** – плотность вероятности нормального распределения;
- **pnorm (x,m,o)** – функция нормального распределения;
- **qnorm (P,m,o)** – обратная функция нормального распределения;
- **rnorm(M,m,o)** – вектор M независимых случайных чисел, каждое из которых имеет нормальное распределение;
  - *x* – значение случайной величины;
  - *P* – значение вероятности;
  - *m*– математическое ожидание;
  - *o* – среднеквадратичное отклонение.

Математическое ожидание и дисперсия являются, по сути, параметрами распределения. Плотность распределения для трех пар значений параметров показана на рис. 1. Напомним, что плотность распределения **dnorm** задает вероятность попадания случайной величины *x* в малый интервал от *x* до *x+dx*. Таким образом, например, для первого графика (сплошная линия) вероятность того, что случайная величина *x* примет значение в окрестности нуля, приблизительно в три раза больше, чем вероятность того, что она примет значение в окрестности *x=2*. А значения случайной величины, большие 5 и меньшие -5, и вовсе маловероятны.

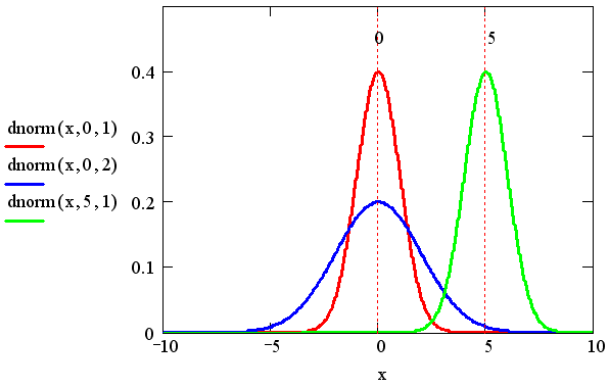

Рис. 1. Плотность вероятности нормальных распределений

Функция распределения F(X) (**cumulative probability**) – это вероятность того, что случайная величина примет значение меньшее или равное

$F(x) = P(X \prec x)$. Как следует из математического смысла, она является интегралом от плотности вероятности в пределах от – *x* до *x*. Функции распределения для упомянутых нормальных законов изображены на рис. 2. Функция, обратная *F(X)* (**inverse cumulative probability**), называемая еще квантилем распределения, позволяет по заданному аргументу *p* определить значение *x*, причем случайная величина будет меньше или равна *x* с вероятностью *p*.

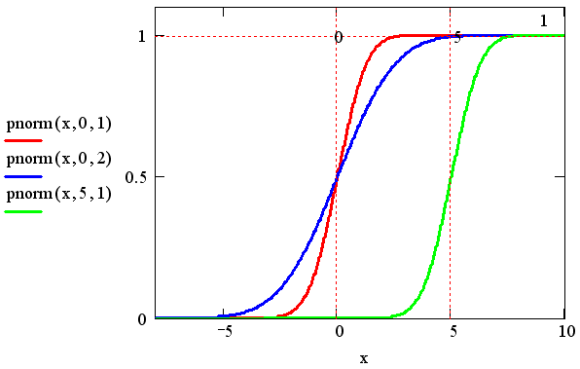

Рис. 2. Нормальные функции распределения

Здесь и далее графики различных статистических функций, показанные на рисунках, получены с помощью Mathcad без каких-либо дополнительных выражений в рабочей области.

Приведем несколько примеров, позволяющих почувствовать математический смысл рассмотренных функций на примере случайной величины *x*, распределенной по нормальному закону с *m*=0 и *o*=1 (листинги 1-5).

Листинг 1. Вероятность того, что x будет меньше 1.881:

$$\text{pnorm}(1.881, 0, 1) = 0.97$$

Листинг 2. 97%-ный квантиль нормального распределения:

$$\text{qnorm}(0.97, 0, 1) = 1.881$$

Листинг 3. Вероятность того, что x будет больше 2:

$$1 - \text{pnorm}(2, 0, 1) = 0.023$$

Листинг 4. Вероятность того, что ж будет находиться в интервале (2,3):

$$\text{pnorm}(3, 0, 1) - \text{pnorm}(2, 0, 1) = 0.021$$

$$\frac{1}{2} \cdot \left(\text{erf}\left(\frac{3}{\sqrt{2}} \right) - \text{erf}\left(\frac{2}{\sqrt{2}} \right) \right) = 0.021$$

Листинг 5. Вероятность того, что |x|<2:

$$\text{pnorm}(2,0,1) - \text{pnorm}(-2,0,1) = 0.954$$

$$\text{erf}\left(\frac{2}{\sqrt{2}}\right) = 0.954$$

Обратите внимание, что задачи двух последних листингов решаются двумя разными способами. Второй из них связан с еще одной встроенной функцией **erf**, называемой функцией ошибок (или интегралом вероятности, или функцией Крампа).

- erf (x) – функция ошибок;
- erfc(x)=1-erf(x).

Математический смысл функции ошибок пояснен из листинга 5. Интеграл вероятности имеет всего один аргумент, в отличии от функции нормального распределения. Исторически, последняя пересчитывалась через табулированный интеграл вероятности по формулам, приведенным в листинге 6 для произвольных значений параметров *m* и *o* (листинг 6).

Листинг 6. Вероятность того, что *x* будет в интервале (2,3):

$$\mu := 5 \qquad \sigma := 2$$

$$\text{pnorm}(3,\mu,\sigma) - \text{pnorm}(2,\mu,\sigma) = 0.092$$

$$\frac{1}{2}\cdot\left(\text{erf}\left(\frac{3-\mu}{\sigma\cdot\sqrt{2}}\right) - \text{erf}\left(\frac{2-\mu}{\sigma\cdot\sqrt{2}}\right)\right) = 0.092$$

Если Вы имеете дело с моделированием методами Монте-Карло, то в качестве генератора случайных чисел с нормальным законом распределения применяйте встроенную функцию **rnorm**. В листинге 7 ее действие показано на примере создания двух векторов по M=500 элементов в каждом с независимыми псевдослучайными числами x_1 и x_2 распределенными согласно нормальному закону.

Листинг 7. Генерация двух векторов с нормальным законом распределения:

$$\sigma := 1 \qquad \mu := 0$$

$$M := 500$$

$$x1 := \text{rnorm}(M,\mu,\sigma)$$

$$x2 := \text{rnorm}(M,\mu,\sigma)$$

О характере распределения случайных элементов векторов можно судить по рис. 3. В дальнейшем мы будем часто сталкиваться с генерацией случайных чисел и расчетом их различных средних характеристик.

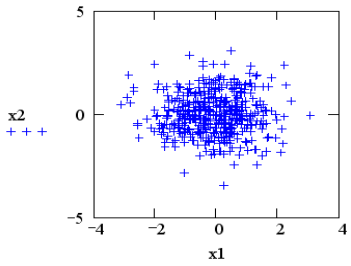

Рис. 3. Псевдослучайные числа с нормальным законом распределения (листинг 7)

Равномерное распределение

Самое простое распределение случайной величины – это распределение с постоянной плотностью вероятности *W(x)*. $W(x)=\frac{1}{(b-a)}$ при $x\in[a,b]$ и *W(x)* =0, для *x* вне интервала $[a,b]$. Эту плотность вероятности, наряду с прочими статистическими характеристиками, задают следующие встроенные функции:

- **dunif (x,a,b)** – плотность вероятности равномерного распределения;
- **punif(x,a,b)** – функция равномерного распределения;
- **qunif(p,a,b)** – квантиль равномерного распределения;
- **runif (м,a,b)** – вектор м независимых случайных чисел, каждое из которых имеет равномерное распределение;
- **rnd (x)** – случайное число, имеющее равномерную плотность распределения на интервале (о, х);
  - х – значение случайной величины;
  - Р – значение вероятности;
  - (a,b) – интервал, на котором случайная величина распределена равномерно.

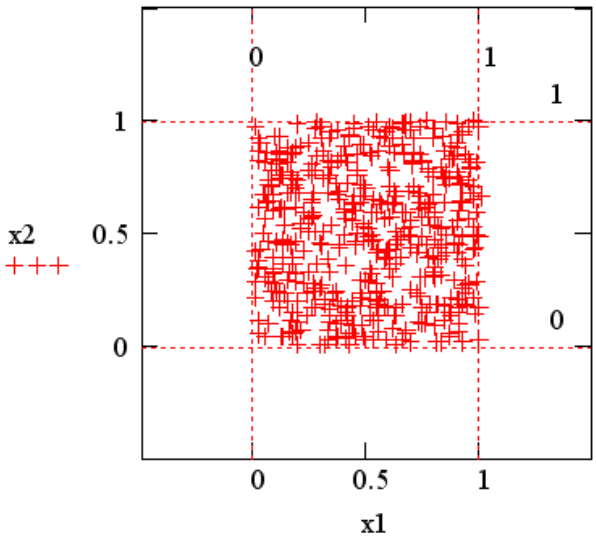

Рис. 4. Псевдослучайные числа с равномерным законом распределения

Чаще всего в несложных программах применяется последняя функция, которая приводит к генерации одного псевдослучайного числа. Наличие такой встроенной функции в Mathcad – дань традиции, применяемой в

большинстве сред программирования. Пример использования генератора вектора из *м* случайных чисел показан на рис. 4, который получен заменой в двух последних строках листинга 7 генератора нормальных чисел на **runif (м,0, 1).** Плотность вероятности и функция равномерного распределения показаны на рис. 5.

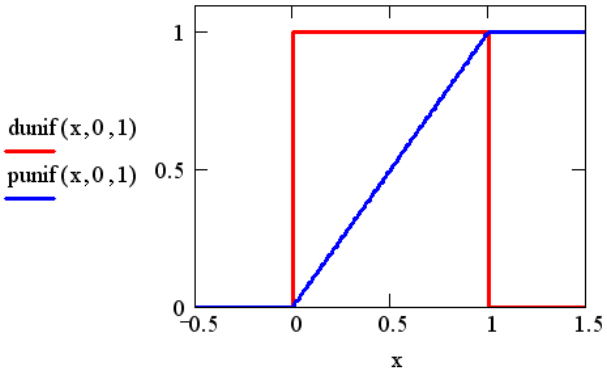

Рис. 5. Плотность вероятности и функция равномерного распределения

Другие статистические распределения

Как легко заметить по рассмотренным трем распределениям, Mathcad имеет четыре основные категории встроенных функций. Они различаются написанием их первой литеры, а оставшаяся часть имени функций (ниже в списке функций она условно обозначена звездочкой) идентифицирует тот или иной тип распределения.

- **d*(x,par)** – плотность вероятности;
- **p*(x,par)** – функция распределения;
- **q*(p,par)** – квантиль распределения;
- **r* (м,par)** – вектор м независимых случайных чисел, каждое из которых имеет соответствующее распределение;
  - *x* – значение случайной величины (аргумент функции);
  - *P* – значение вероятности;
  - *par* – список параметров распределения.

Чтобы получить функции, относящиеся, например, к равномерному распределению, вместо * надо поставить **unif** и ввести соответствующий список параметров **par**. Он будет состоять в данном случае из двух чисел *a,b* – интервала распределения случайной величины.

Перечислим все типы распределения, реализованные в Mathcad, вместе с их параметрами, на этот раз обозначив звездочкой * недостающую первую букву встроенных функций. Некоторые из плотностей вероятности показаны на рис. 7.

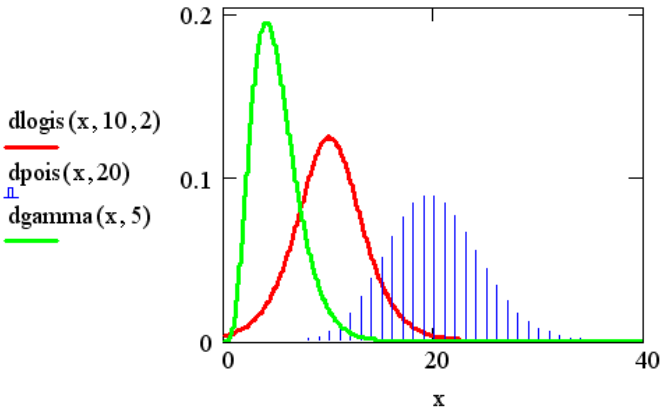

Рис. 7. Плотность вероятности некоторых распределений

- ***beta(x,s_1,s_2)** – бета-распределение (s_1,s_2 параметры принимающие только положительные значения, 0<x<1).
- ***binom(k,n,p)** – биномиальное распределение (*n* – целый параметр, *0<k<n* и *0<p<1* – параметр, равный вероятности успеха единичного испытания).
- ***cauchy(x,l,s)** – распределение Коши (l – параметр разложения, s>0 – параметр масштаба).
- ***chisq(x,d)** – x^2 ("хи-квадрат") распределение (d>0 – число степеней свободы).
- ***exp(x,r)** – экспоненциальное распределение (r>0 – показатель экспоненты).
- ***F(x,d_1,d_2)** – распределение Фишера (d_1,d_2>0 – числа степеней свободы).
- ***gamma(x,s)** – гамма-распределение (s>0 – параметр формы).
- ***geom(k,p)** – геометрическое распределение (0<p<1 – параметр, равный вероятности успеха единичного испытания).
- ***hypergeom(k,a,b,n)** – гипергеометрическое распределение (a,b,n – целые параметры).
- ***lnorm(x,m,o)** – логарифмически нормальное распределение (m – натуральный логарифм математического ожидания, o>0 – натуральный логарифм среднеквадратичного отклонения).
- ***logis(x,l,s)** – логистическое распределение (l – математическое ожидание, s>0 – параметр масштаба).
- ***nbinom(k,n,p)** – отрицательное биномиальное распределение (n>0 – целый параметр, 0<p<1).
- ***norm(x,m,o)** – нормальное распределение (m– среднее значение, o>0 – среднеквадратичное отклонение).
- ***pois(k,a)** – распределение Пуассона (a>0 – параметр).
- ***t(x,d)** – распределение Стьюдента (d>0 – число степеней свободы).
- ***unif(x,a,b)** – равномерное распределение (a<b – границы интервала).
- ***weibuli(x,s)** – распределение Вейбулла (s>0 – параметр).

Вставку рассмотренных статистических функций в программы удобно осуществлять с помощью диалогового окна **Insert Function** (Вставка функции). Для этого необходимо выполнить следующие действия:

- Установите курсор на место вставки функции в документе.
- Вызовите диалоговое окно **Insert Function** нажатием кнопки f(x) на стандартной панели инструментов или командой меню **Insert › Function** (Вставка › Функция), или нажатием клавиш **CTRL + E**.

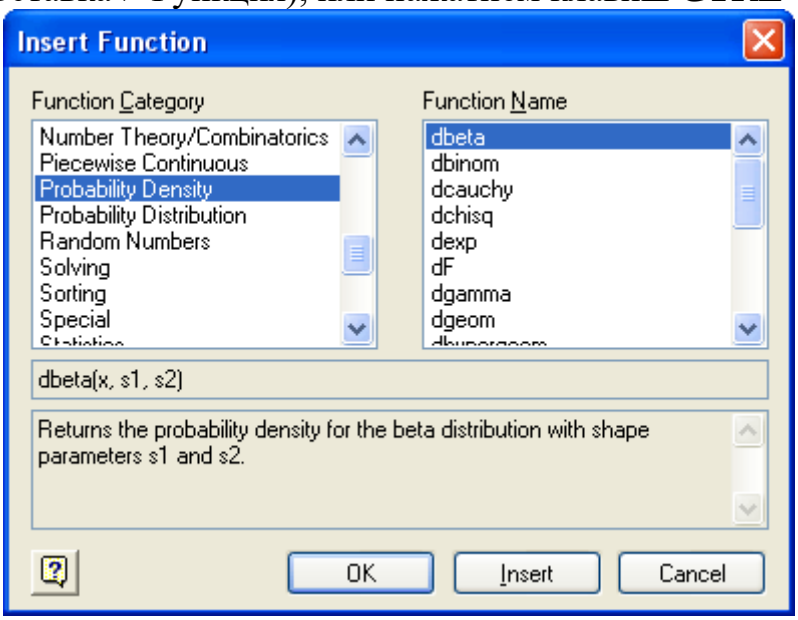

Рис. 8. Диалоговое окно Insert Function

- В списке **Function Category** (Категория функции) (рис.8) выберите одну из категорий статистических функций. Категория **Probability Density** (Плотность вероятности) содержит встроенные функции для плотности вероятности, **Probability Distribution** (Функция распределения) – для вставки функций или квантилей распределения, **Random Numbers** (Случайные числа) – для вставки функции генерации случайных чисел.
- В списке **Function Name** (Имя функции) выберите функцию, в зависимости от требующегося закона распределения. При выборе того или иного элемента списка в текстовых полях в нижней части окна будет появляться информация о назначении выбранной функции.
- Нажмите кнопку **ОК** для вставки функции в документ.

2. Статистические характеристики. Построение гистограмм.

В большинстве статистических расчетов Вы имеете дело либо со случайными данными, полученными в ходе какого-либо эксперимента (которые выводятся из файла или печатаются непосредственно в документе), либо с результатами генерации случайных чисел, рассмотренными в предыдущих разделах встроенными функциями, моделирующими то или иное явление методом Монте-Карло. Рассмотрим возможности Mathcad по оценке функций распределения и расчету числовых характеристик случайных данных.

В Mathcad имеется ряд встроенных функций для расчетов числовых статистических характеристик рядов случайных данных.

- mean(x) – выборочное среднее значение;
- median (x) – выборочная медиана (**median**) – значение аргумента, которое делит гистограмму плотности вероятностей на две равные части;
- var(x) – выборочная дисперсия (**variance**);
- stdev(x) – среднеквадратичное (или "стандартное") отклонение (**standard deviation**);
- max(x), mm (x) – максимальное и минимальное значения выборки;
- mode(x) – наиболее часто встречающееся значение выборки;
- var (x),stdev(x) – выборочная дисперсия и среднеквадратичное отклонение в другой нормировке;
  - x – вектор (или матрица) с выборкой случайных данных.

Пример использования первых четырех функций приведен в листинге 8.1.

Листинг 8.1. Расчет числовых характеристик случайного вектора:

$x := \text{rweibull}(1000, 1.5)$

$N := \text{length}(x) \qquad N = 1 \times 10^3$

$\text{mean}(x) = 0.917$

$\text{median}(x) = 0.781$

$\text{var}(x) = 0.402$

$\text{stdev}(x) = 0.634 \qquad \sqrt{\text{var}(x)} = 0.634$

$hi := \text{mean}(x) + \text{stdev}(x) \qquad lo := \text{mean}(x) - \text{stdev}(x)$

Гистограммой называется график, аппроксимирующий по случайным данным плотность их распределения. При построении гистограммы область значений случайной величины **(a,b)** разбивается на некоторое количество сегментов - **bin**, а затем подсчитывается процент попадания данных в каждый сегмент. Для построения гистограмм в Mathcad имеется несколько встроенных функций. Рассмотрим их, начиная с самой сложной по применению, чтобы лучше разобраться в возможностях каждой из функций.

Гистограмма с произвольными сегментами разбиения

- **hist(intvis,x)** – вектор частоты попадания данных в интервалы гистограммы;
  - ***intvis*** – вектор, элементы которого задают сегменты построения гистограммы в порядке возрастания $a<intvis_i<b$;
  - ***x*** – вектор случайных данных.

Если вектор **intvis** имеет **bin** элементов, то и результат **hist** имеет столько же элементов. Построение гистограммы иллюстрируется листингом 8.2 и рис. 9.

Листинг 8.2. Построение гистограммы:

$$N := 1000$$

$$bin := 30$$

$$x := rnorm(N, 0, 1)$$

$$lower := floor(min(x))$$

$$upper := ceil(max(x))$$

$$h := \frac{upper - lower}{bin}$$

$$j := 0 .. bin$$

$$int_j := lower + h \cdot j$$

$$f := \frac{1}{N \cdot h} \cdot hist(int, x)$$

Для анализа взято *N=1000* данных с нормальным законом распределения, созданных генератором случайных чисел (третья строка листинга). Далее определяются границы интервала (**upper**, **lower**), содержащего внутри себя все случайные значения, и осуществляется его разбиение на количество (**bin**) одинаковых сегментов, начальные точки которых записываются в вектор **int** (предпоследняя строка листинга).

В векторе **int** можно задать произвольные границы сегментов разбиения так, чтобы они имели разную ширину.

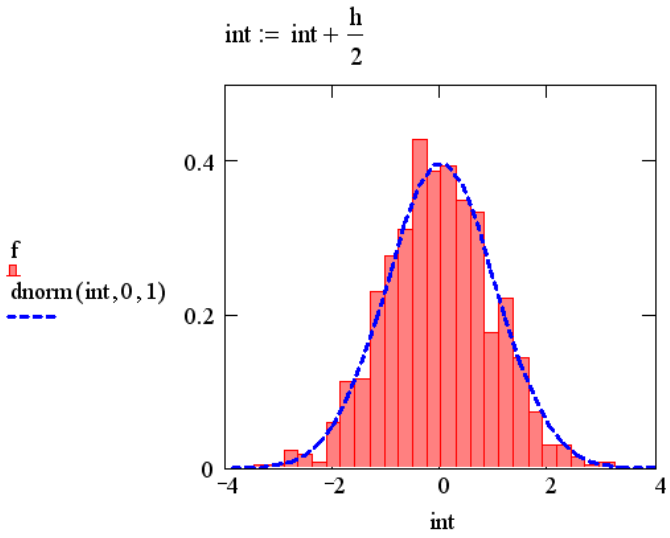

Рис. 9. Построение гистограммы (листинг 8)

Обратите внимание, что в последней строке листинга осуществлена нормировка значений гистограммы, с тем чтобы она правильно аппроксимировала плотность вероятности, также показанную на графике. Очень важно переопределение вектора **int** в самом верху рис. 9, которое необходимо для перехода от левой границы каждого элементарного сегмента к его центру.

Гистограмма с разбиением на равные сегменты

Если нет необходимости задавать сегменты гистограммы разной ширины, то удобнее воспользоваться упрощенным вариантом функции **hist**.

- **hist (bin, x)** – вектор частоты попадания данных в интервалы гистограммы;
  - **bin** – количество сегментов построения гистограммы;
  - **x** – вектор случайных данных.

Для того чтобы использовать этот вариант функции **hist** вместо предыдущего, достаточно заменить первый из ее аргументов в листинге 8.

Недостаток упрощенной формы функции hist в том, что по-прежнему необходимо дополнительно определять вектор сегментов построения гистограммы. От этого недостатка свободна появившаяся в Mathcad 14 функция **histogram.**

- **histogram (bin, x)** – матрица гистограммы размера **bin** x **2**, состоящая из столбца сегментов разбиения и столбца частоты попадания в них данных;
  - **bin** – количество сегментов построения гистограммы;
  - **x** – вектор случайных данных.

Примеры использования функции **histogram** приведены в листинге 9 и рис. 10.

Листинг 9 Применение функции histogram

нормальное распределение V := rnorm(1000,10,1)

построение гистограммы

nn := 21 hh := histogram(nn,V)

int := $hh^{\langle 1 \rangle}$ вектор средних точек интервалов

h := $hh^{\langle 2 \rangle}$ частота попадания чисел в интервал

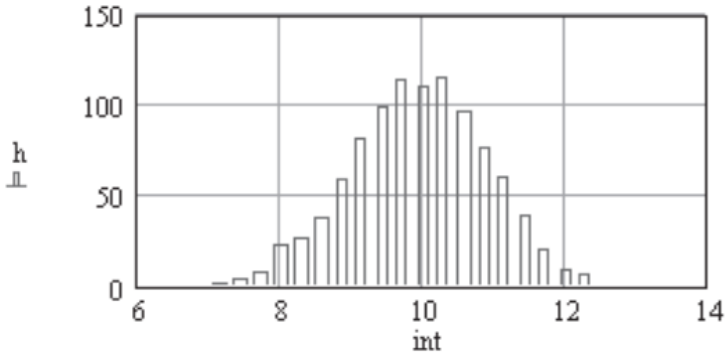

Рис. 10. График гистограммы (листинг9)

Создание графика гистограммы

Для того чтобы создать график в виде гистограммы:

- Постройте двумерный график, задайте переменные по осям и пределы оси *x* (числа **lower** и **upper**).
- Войдите в диалоговое окно **Formatting Currently Selected Graph** (Форматирование) выбранного графика (например, двойным щелчком мыши) и перейдите на вкладку **Traces** (Графики).
- Установите для серии данных гистограммы в поле **Type** (Тип) элемент списка **bar** (столбцы) или **solidbar** (гистограмма) (рис. 11).
- Нажмите кнопку **ОК**.

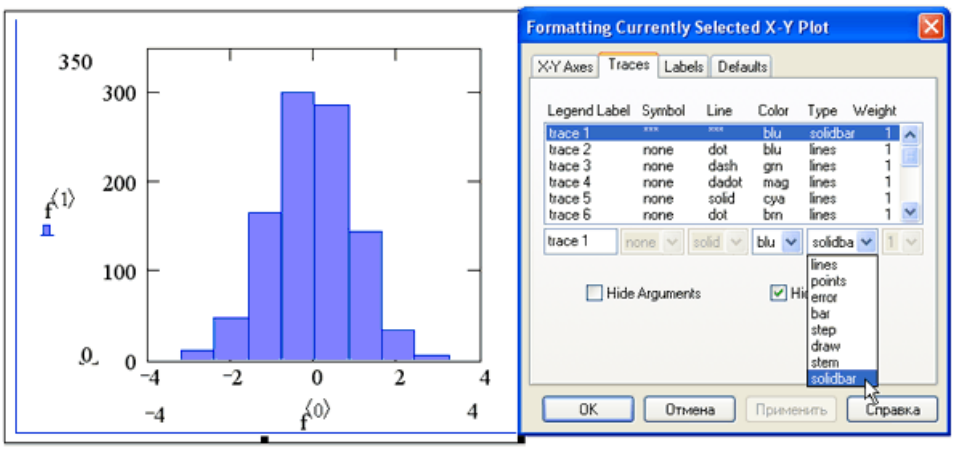

Рис. 11. Установка типа графика для построения гистограммы

3.Случайные процессы

Встроенные функции для генерации случайных чисел создают выборку из случайных данных. Часто требуется создать непрерывную или дискретную случайную функцию *A(t)* одной или нескольких переменных (случайный процесс или случайное поле), значения которой будут упорядочены относительно своих переменных. Создать псевдослучайный процесс можно способом, представленным в листинге 10.

Листинг 10. Генерация псевдослучайного процесса:

$$N := 20$$

$$\tau := 0.5$$

$$Tmax := (N - 1)\cdot\tau$$

$$j := 0 .. N - 1$$

$$T_j := j\cdot\tau$$

$$x := rnorm(N, 0, 1)$$

$$KS1 := cspline(T, x)$$

$$A(t) := interp(KS1, T, x, t)$$

В первой строке листинга 10 определено количество N независимых случайных чисел, которые будут впоследствии сгенерированы, и радиус временной корреляции τ. В следующих трех строках определяются моменты времени T_j, которым будут отвечать случайные значения A(t). Создание нормального случайного процесса сводится к генерации обычным способом вектора независимых случайных чисел x и построению интерполяционной зависимости в промежутках между ними. В листинге 10 используется сплайн-интерполяция.

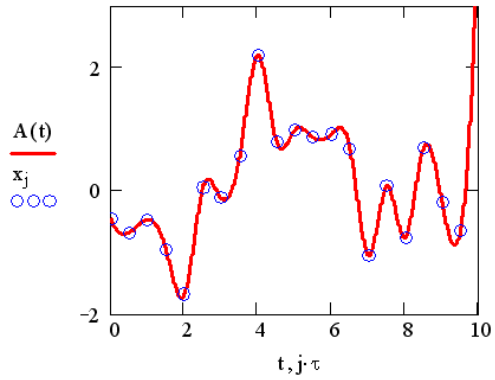

Рис. 12. Псевдослучайный процесс (листинг 10)

В результате получается случайный процесс *A(t),* радиус корреляции которого определяется расстоянием τ между точками, для которых строится интерполяция. График случайного процесса *A(t)* вместе с исходными случайными числами показан на рис. 12. Случайное поле можно создать несколько более сложным способом с помощью многомерной интерполяции. Вопросы обработки данных рассмотрены в следующей лекции.