Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

УДК 519.637

# СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ В СРЕДЕ MATHCAD

*С.В. Шушкевич, Г.Ч. Шушкевич*

*В статье излагаются основные методы статистической обработки данных с использованием встроенных функций системы MathCad. Рассмотрены встроенные функции для вычисления числовых характеристик выборки, коэффициента корреляции по Пирсону, построения регрессий, описываемых функциями разного вида, выполнения сглаживания эмпирических данных.*

## Введение

В настоящее время информация становится частью действительности, требуя адекватных технологий ее восприятия и анализа. Существует множество отраслей знаний, в которых возникает необходимость проведения анализа данных – медицина, психология, биология, социология, экономика и др. Статистический анализ позволяет компактно описать данные, понять их структуру, провести классификацию, увидеть закономерности в случайных процессах.

## Описательная статистика

Описательная статистика отражает числовые характеристики данных, полученных при наблюдении или в эксперименте. В системе Mathcad в разделе Statistics представлены встроенные функции для вычисления числовых характеристик выборочных данных, размещенных в матрице A размерности $n \times m$:

- mode(A) – возвращает моду выборки – наиболее часто встречающееся значение выборочных данных;
- median(A) – возвращает значение, стоящее в середине выборки. Если число элементов выборки четное $n = 2k$, то медиана определяется по формуле $(x_k + x_{k+1})/2$. При нечетном объеме выборки $n = 2k+1$ значение медианы равно $x_{k+1}$;
- mean(A) – возвращает среднее значение выборки, вычисленное по формуле $\frac{1}{n\,m}\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}$,
- hmean(A) – возвращает среднее гармоническое значение выборки, вычисленное по формуле $\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{a_{ij}}\right)^{-1}$, $a_{ij} > 0$,

Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

- gmean(A) – возвращает среднее геометрическое значение выборки, вычисленное по формуле $\left(\prod_{i=1}^{n}\prod_{j=1}^{m} a_{ij}\right)^{1/mn}$, $a_{ij} > 0$,
- var(A) – возвращает значение выборочной дисперсии, вычисляемой по формуле $\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij} - \text{mean}(A))^2$;
- Var(A) – возвращает значение исправленной выборочной дисперсии, вычисляемой по формуле $\frac{nm}{nm-1}\text{var}(A)$;
- stdev(A) – возвращает значение среднеквадратичного отклонения, которое вычисляется по формуле $\sqrt{\text{var}(A)}$;
- Stdev(A) – возвращает значение исправленного среднеквадратичного отклонения, которое вычисляется по формуле $\sqrt{\text{Var}(A)}$;
- kurt(A) – возвращает значение эксцесса выборки. Он характеризует высоковершинность или низковершинность экспериментального распределения по отношению к нормальному распределению. Если эксцесс больше нуля, то распределение имеет более острую вершину, чем распределение Гаусса, если меньше нуля – то более плоскую. Эксцесс вычисляют по формуле

$$\frac{nm(nm+1)}{(nm-1)(nm-2)(nm-3)}\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij} - \text{mean}(A))^4\,\text{Var}(A)^{-2} - \frac{3(nm-1)^2}{(nm-2)(nm-3)};$$

- skew(A) – возвращает значение асимметрии выборки. Асимметрия характеризует скошенность эмпирического распределения по отношению к нормальному. Если асимметрия больше нуля, то имеем левостороннюю асимметрию, если меньше нуля – то правостороннюю. Асимметрию вычисляют по формуле

$$\frac{nm}{(nm-1)(nm-2)}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{a_{ij} - \text{mean}(A)}{\text{Stdev}(A)}\right)^3.$$

Для приведенных встроенных функций аргументами могут быть любое количество векторов, матриц и чисел.

Гистограмма аппроксимирует плотность эмпирического распределения по выборочным данным. При построении гистограммы область наблюдаемых значений случайной величины разбивается на n равных интервалов и подсчитывается количество попаданий данных в каждый интервал. Для построения гистограммы используют функцию histogram (n, A), которая возвращает матрицу размера $n \times 2$, состоящую из столбца интервалов разбиения и столбца количества данных в каждом их них.

Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

*Пример 1.* Найти числовые характеристики выборки. Построить гистограмму относительных частот с 7 интервалами и график плотности нормального распределения с математическим ожиданием, равным выборочному среднему, и среднеквадратичным отклонением, равным выборочному среднеквадратичному отклонению.

*Math – Документ*

1. Ввод данных:

dt := D:\ПОС...\dt.xls dt =

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4.622 | 1.259 | 0.053 | 1.903 | 0.272 | 1.087 | 0.759 | 2.113 | 3.384 | 2.617 |
| 2.971 | 2.724 | 2.831 | 2.346 | 1.089 | 1.138 | 0.511 | 2.393 | 0.636 | 1.289 |
| 0.446 | 2.033 | 2.116 | 1.151 | 1.179 | 3.824 | 0.411 | 1.604 | 3.215 | 3.785 |
| 1.802 | 1.37 | 3.912 | 1.923 | 1.395 | 4.198 | 2.542 | 1.612 | 1.023 | 0.529 |
| 1.982 | 0.348 | 0.736 | 3.036 | 1.37 | 2.027 | 2.48 | 0.967 | 1.558 | 1.324 |

2. Вычисление числовых характеристик:

mode(dt) = 1.37 median(dt) = 1.608

mean(dt) = 1.838 hmean(dt) = 0.829 gmean(dt) = 1.434

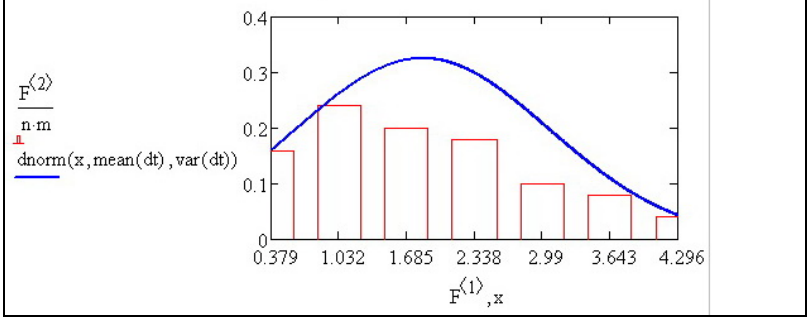
var(dt) = 1.225 Var(dt) = 1.25

stdev(dt) = 1.107 Stdev(dt) = 1.118 kurt(dt) = −0.275 skew(dt) = 0.613

3. Построение гистограммы и графика плотности нормального распределения:

F := histogram(7, dt)

$F^T$ = (0.379 1.032 1.685 2.338 2.99 3.643 4.296 / 8 12 10 9 5 4 2)




**Корреляционный и регрессионный анализ**

Пусть наблюдаются два признака X, У, значения которых являются случайными величинами и замерены на одной выборке. Между ними существует связь особого рода, при которой изменение значений одной случайной величины ведет к изменению распределения значений другой. Такая связь называется стохастической [1]. Для оценки силы стохастической связи используют коэффициент корреляции. Прямолинейная зависимость, как правило, наблюдается между величинами, имеющими нормальное

Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

распределение. В этом случае для вычисления коэффициента корреляции Пирсона r по формуле

$$\frac{\mathrm{cvar}(X,Y)}{\mathrm{stdev}(X)\,\mathrm{stdev}(Y)}$$

используют встроенную функцию corr(X,У), здесь встроенная функция cvar(X,У) вычисляет ковариацию элементов массивов X, У по формуле

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mathrm{mean}(X))(y_i-\mathrm{mean}(Y)).$$

Уравнение регрессии представляет собой функцию, аппроксимирующую зависимость между случайными величинами X и Y. По форме зависимости различают:

1) линейную регрессию, которая выражается линейной функцией $y=ax+b$,

2) нелинейную регрессию, которая выражается полиномиальной, степенной, логарифмической, тригонометрическими и другими функциями.

Все встроенные функции, относящие к построению различных уравнений регрессии, находятся в разделе Curve Fitting and Smoothing.

Для нахождения коэффициентов линейной функции можно воспользоваться встроенными функциями:

- line (vx,vy) – возвращает вектор из двух элементов (b,a),
- slope (vx, vy) – возвращает значение коэффициента a,
- intercept (vx, vy) – возвращает значение коэффициента b,

где vx, vy – вектора наблюдаемых данных одинакового размера.

Встроенная функция medfit(vx,vy) предназначена для нахождения коэффициентов уравнения прямой вида $y=ax+b$, наилучшим образом аппроксимирующей данные векторов vx и vy с использованием медиан-медианной регрессии – возвращает вектор из двух элементов (b, a).

Встроенная функция stderr(vx,vy) возвращает среднеквадратичную ошибку, связанную с линейной регрессией для точек, описанных векторами vx и vy, – измеряет разброс данных относительно линии регрессии:

$$\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(vy_i-(\mathrm{interpt}(vx,vy)+\mathrm{slope}(vx,vy)vx_i)\right)^2}$$

Построение полиномиальной регрессии $P_n(t)=a_0+a_1t+\ldots+a_nt^n$ осуществляется с помощью комбинации двух встроенных функций:

- regress(vx,vy,n) – возвращает вектор-столбец значений коэффициентов $a_i,\ i=0,1,\ldots,n,$ для построения полинома $P_n(t)=a_0+a_1t+\ldots+a_nt^n$,
- interp(regress(vx,vy.к),vx,vy,t) – возвращает значение полинома $P_n(t)$ в точке t.

Для построения полинома n-ой степени необходима, по крайней мере, (n+1) точка исходных данных. На практике не рекомендуется использовать

Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

полином выше четвертой – шестой степени, поскольку погрешности реализации регрессии сильно возрастают.

Полиномиальную регрессию можно вычислить также путем сшивания нескольких полиномов второго порядка $P_2(t)$:

- loess (vx,vy,s) – вектор коэффициентов для построения регрессии сшивкой полиномов $P_2(t)$;
- interp(loess (vx,vy,s),vx,vy,t) – результат полиномиальной регрессии.

Параметр s определяет размер окрестности данных, по умолчанию s=0.75.

Если Вы хорошо представляете, какой зависимостью описываются данные в массиве исходных данных, то для построения уравнений таких регрессии имеются специальные встроенные функции. Для использовании этих встроенных функций необходимо задать вектор начальных значений vn для коэффициентов $a, b, c$ размерности $3\times1$. Каждая из нижеприведенных функций возвращает вектор уточненных значений параметров $a, b, c$ для конкретной кривой:

- expfit(vx,vy,vn) – для функции вида $f(t) = ae^{bt} + c$,
- lgsfit(vx,vy,vn) – для функции вида $f(t) = a / (1 + be^{-ct})$,
- sinfit(vx,vy,vn) – для функции вида $f(t) = a\sin(x + b) + c$,
- pwfit(vx,vy,vn) – для функции вида $f(t) = ax^b + c$,
- logfit (vx,vy,vn) – для функции вида $f(t) = a\ln(x + b) + c$,
- lnfit(x,y) – для функции вида $f(t) = a\ln(x) + b$.

Затем по уточным значениям $a, b, c$ строится линия регрессии.

В системе представлена встроенная функция для нахождения регрессии в виде линейной комбинации $a_0f_0(t) + a_1f_1(t) + \ldots + a_nf_n(t)$, где $f_i(t)$ – известные функции пользователя, коэффициенты $a_i$ подлежат определению с помощью встроенной функции linfit(vx,vy,F), где F(t) – вектор-столбец, содержащий функции $f_i(t)$.

Можно также найти регрессии в виде $a_0f_0(a_1, t) + a_2f_1(a_3, t) + \ldots$, где $f_i(a_{2i+1}, t)$ – известные функции пользователя с неизвестными коэффициентами, коэффициенты $a_i$ подлежат определению с помощью встроенной функции genfit(vx,vy,a,G), здесь a – вектор начальных значений для коэффициентов $a_i$, G(t,a) – вектор–столбец, составленный из функции пользователя и ее производных по $a_i$.

По аналогии с одномерной полиномиальной регрессией можно выполнить построение двумерной полиномиальной регрессии. Для этого используют те же встроенные функции, что и для одномерной регрессии:

- regress(vxy,vz,n) – возвращает вектор коэффициентов для построения полиномиальной регрессии данных,

Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

- loess(vxy,vz,s) – возвращает вектор коэффициентов для построения полиномиальной регрессии данных кусочными полиномами,
- interp(vr,vxy,vz,$\begin{pmatrix} x \\ y \end{pmatrix}$) – возвращает значение полинома $P_n(x,y)$ в точке $(x,y)$,

где vxy – матрица размерности $m \times 2$, определяющая координаты $m$ точек $(x_i, y_i)$ на плоскости $Oxy$,

vz – вектор значений $z_i$ размерности $m$, соответствующий двумерному массиву vxy на плоскости,

n – степень полинома (целое положительное число),

s – параметр, определяющий размер окрестности данных, по умолчанию s=0.75.

vr – вектор, созданный одной из встроенных функций regress или loess.

*Пример 2*. По выборочным данным построить двумерную полиномиальную регрессию.

*Math – Документ*

1. Исходные данные:

$$vxy := \begin{pmatrix} 2 & 2 & 3 & 1 & 4 & 0 & 3 & 0 & 5 & 4 \\ 2 & 3 & 3 & 4 & 1 & 0 & 1.9 & 5 & 0.5 & 4 \end{pmatrix} \qquad vxy := vxy^T$$

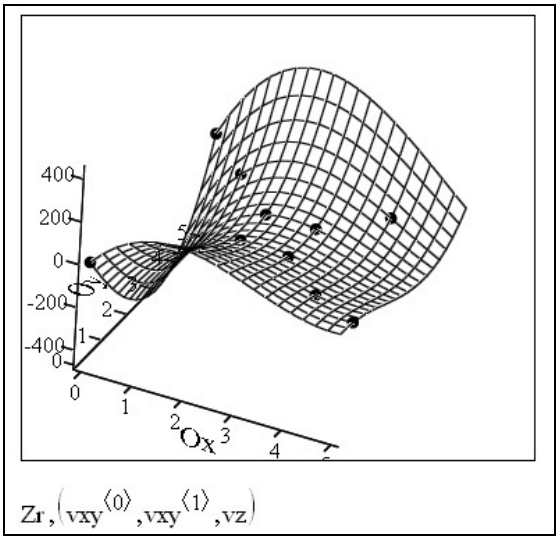
$$vz := (1 \;\; 2 \;\; 2 \;\; 3.5 \;\; 2.3 \;\; 2.6 \;\; 5 \;\; 3 \;\; 2 \;\; 1) \qquad vz := vz^T$$

2. Построение полинома третьей степени с помощью функции regress:

$vr := regress(vxy, vz, 3)$ $vr^T =$

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 0 | 3 | 3 | 3 | -10.991 | 8.278 | 11.35 | 263.603 | ... |

$$Zr(x,y) := interp\left[vr, vxy, vz, \begin{pmatrix} x \\ y \end{pmatrix}\right]$$



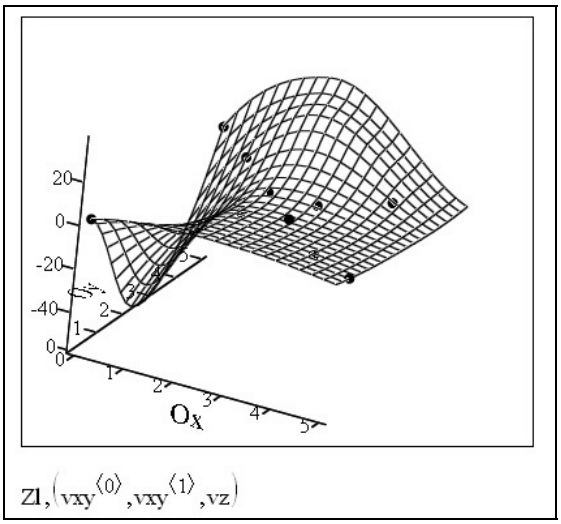
Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

3. Построение полинома с помощью функции loess:

$vr1 := loess(vxy, vz, 1.1)$ $vr1^T =$

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 110 | 2.326 | 4.244 | 5.843 | 3.52 | 3.087 | 2.542 | ... |

$Zl(x,y) := interp\left[vr1, vxy, vz, \begin{pmatrix} x \\ y \end{pmatrix}\right]$




**Сглаживание эмпирических данных**

Вид эмпирической линии регрессии может показать, какая форма связи имеет место в конкретном случае – прямолинейная, параболическая или какая-либо другая [2]. Для получения возможности более точно судить о том, как меняется одна величина, например, Y, при изменении другой, например, X, выполняют элиминирование случайных колебаний эмпирической линии регрессии путем процедуры статистического сглаживания (выравнивания) данных. В системе Mathcad представлены несколько встроенных функций, реализующих различные алгоритмы сглаживания данных – раздел Curve Fitting and Smoothing:

- medsmooth(vy,nn) – возвращает вектор такой размерности, как исходный вектор vy, созданный сглаживанием данных по методу скользящей медианы,
- ksmooth(vx,vy,b) – возвращает вектор такой размерности, как исходный вектор vy, созданный сглаживанием данных на основе функции Гаусса,
- supsmooth(vx,vy) – возвращает вектор сглаженных vy, вычисленных на основе использования процедуры линейного сглаживания методом наименьших квадратов,

где vx – вектор действительных данных аргумента (для supsmooth его элементы должны быть расположены в порядке возрастания);
vy – вектор действительных значений того же размера, что и vx;
nn – нечетное число, меньшее размерности вектора vy,
b – целое число, в несколько раз превышающее интервал между точками $x_i$.

Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

Результаты использования встроенных функций medsmooth, ksmooth, supsmooth для сглаживания экспериментальных данных представлены, соответственно, на рис.1 – рис.3:

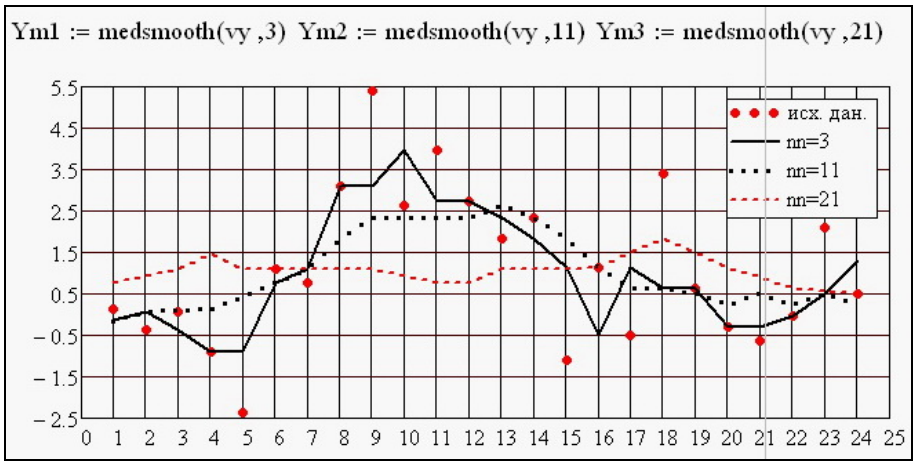



Рис. 1 – Сглаживание данных с помощью medsmooth

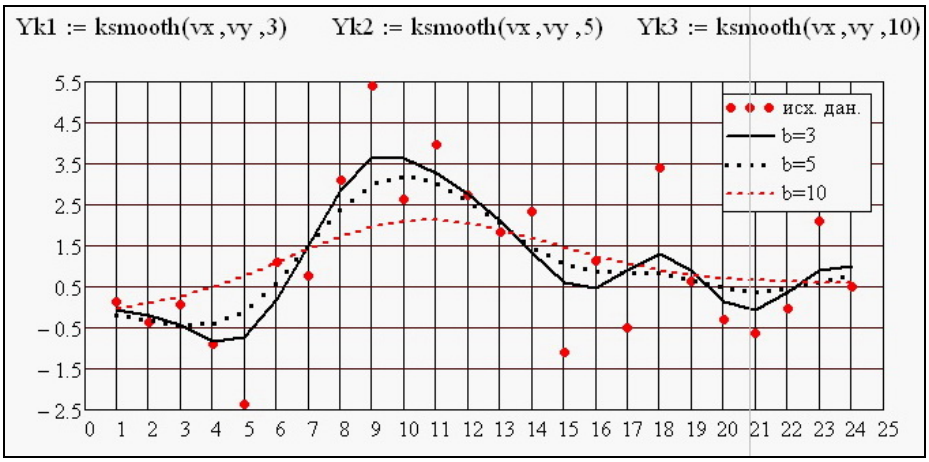



Рис. 2 – Сглаживание данных с помощью ksmooth

Технологии информатизации и управления ТИМ-2011: материалы II Международной научно-практической конференции [Электронный ресурс] / ГУО «ИТИиУ» БГУ. - Гродно, 2011. - 1 электр. компакт диск (CD-R). - 726 с. - Рус. - Деп. в ГУ «БелИСА» 31.08.2011 г., № Д201138

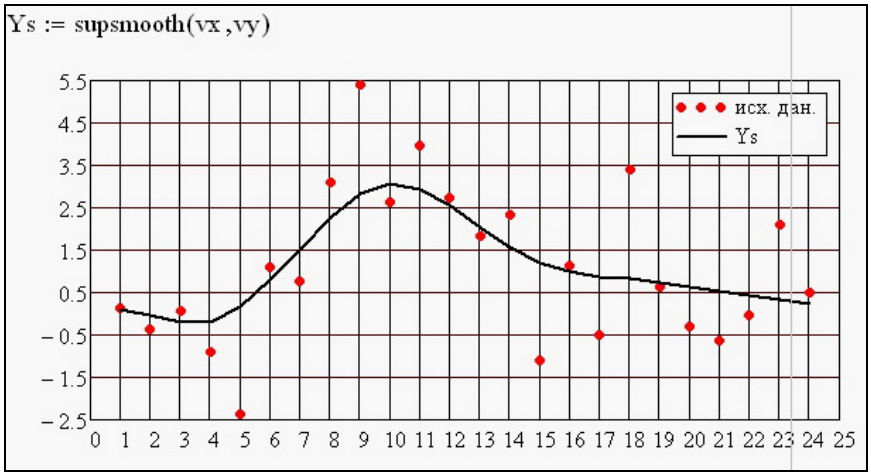



Рис. 3 – Сглаживание данных с помощью supsmooth

## Заключение

Использование систем компьютерной математики позволяет выполнять начальные этапы анализа данных специалистам их конкретных областей. В системе Mathcad можно вычислять простые описательные статистики, строить графики, выполнять группировку данных и строить более сложные статистические модели, например, регрессионные, позволяющие оценивать и прогнозировать поведение и зависимость исследуемых величин.

## Список литературы

1. Мацкевич, И.П. Высшая математика: Теория вероятностей и математическая статистика / И.П. Мацкевич, Г.П. Свирид. – Минск: Выш. шк., 1993. – 269 с.
2. Поллард, Дж. Справочник по вычислительным методам статистики / Дж. Поллард. – М.: Финансы и статистика, 1982. – 343 с.

*Шушкевич Светлана Владимировна, старший преподаватель кафедры экспериментальной и прикладной психологии факультета психологии Гродненского государственного университета имени Янки Купалы, spusha@list.ru*

*Шушкевич Геннадий Чеславович, заведующий кафедрой информатики и компьютерного моделирования факультета математики и информатики Гродненского государственного университета имени Янки Купалы, доктор физико-математических наук, доцент, g_shu@rambler.ru*