Лекция 3

Распределение Гаусса

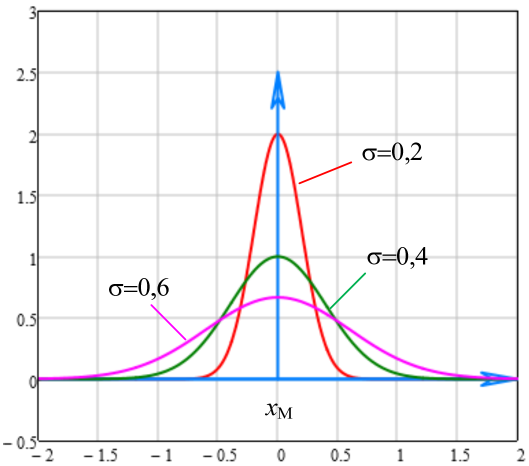
Поскольку Гауссово распределение играет важную роль в технических измерениях, рассмотрим его более подробно.

Гауссова плотность распределение случайной величины x имеет вид

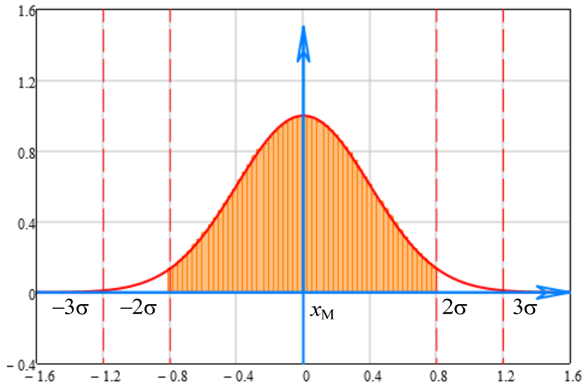
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-x_M)^2}{2\sigma^2}}. \qquad (1)$$

Здесь $\frac{1}{\sigma\sqrt{2\pi}}$ - нормирующий сомножитель такой, что интеграл от плотности вероятности равен единицы. В этой формуле для плотности вероятности имеются две важные величины: x_M - математическое ожидание (появление этой величины мы ожидаем чаще других значений), и дисперсия $\sigma^2 = D$, $\sigma=\sqrt{D}$ - среднеквадратическое отклонение.

Чем хороша эта функция? Она хороша тем, что она полностью определяется своими первыми двумя моментами. То есть для того чтобы восстановить это функцию нужно только математическое ожидание и дисперсия.



На графике видно, что увеличении дисперсии σ плотность вероятности «размазывается». То есть разброс случайной величины увеличивается. Высота каждой кривой определяется соотношением $\sim 1/\sigma\sqrt{2\pi}$



На графике видно, что на диапазоне горизонтального интервала $x\in[-3\sigma,3\sigma]$ практически полностью укладываются все возможные случайной величины с вероятностью 0,997. То же самое можно сказать и об интервале $x\in[-2\sigma,2\sigma]$, с вероятностью 0,95. Эти интервалы называются доверительными.

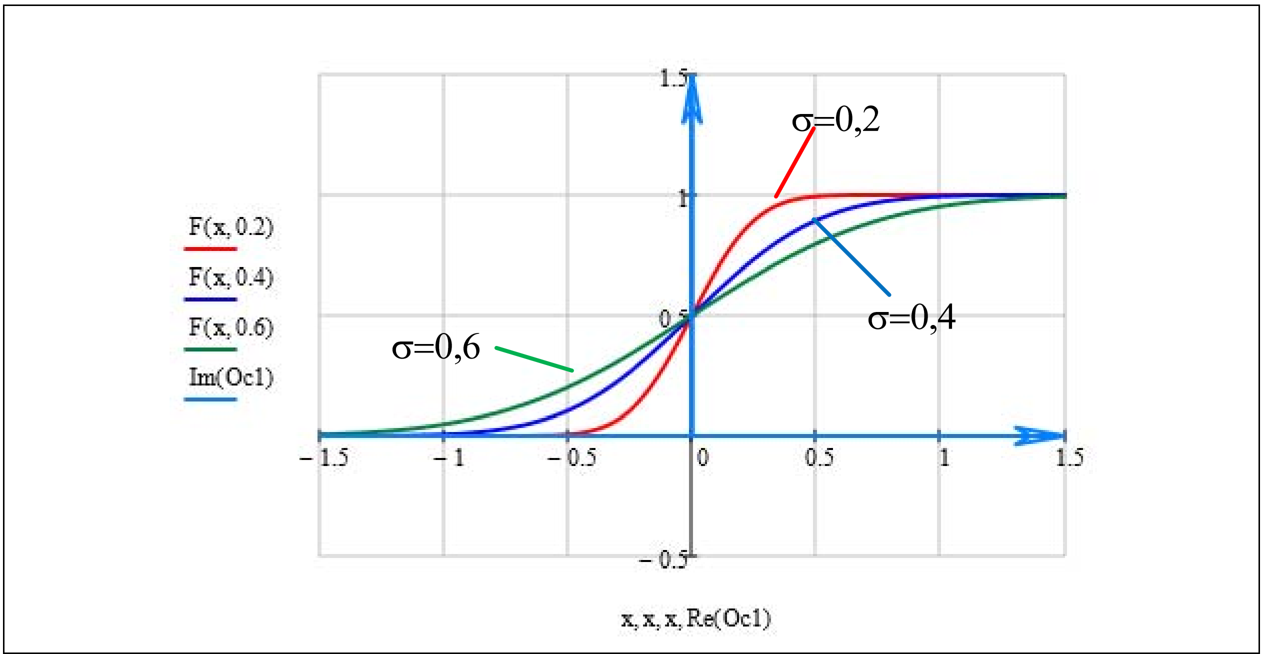
Вероятность появление случайной величины в диапазоне от $-\infty$ до x равна

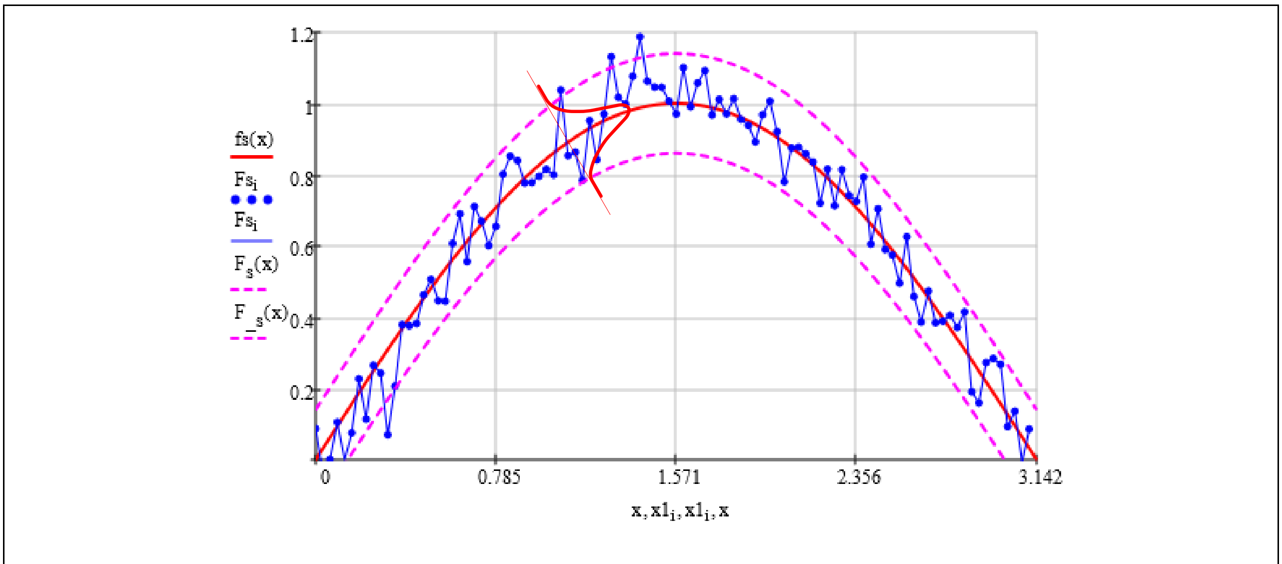
$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(x-x_M)^2}{2\sigma^2}}$$

Зададим диапазон изменения величины в интервале значений $x\in[-3\sigma,3\sigma]$, и в интервале значений $x\in[-2\sigma,2\sigma]$, и получив вероятность этих событий:

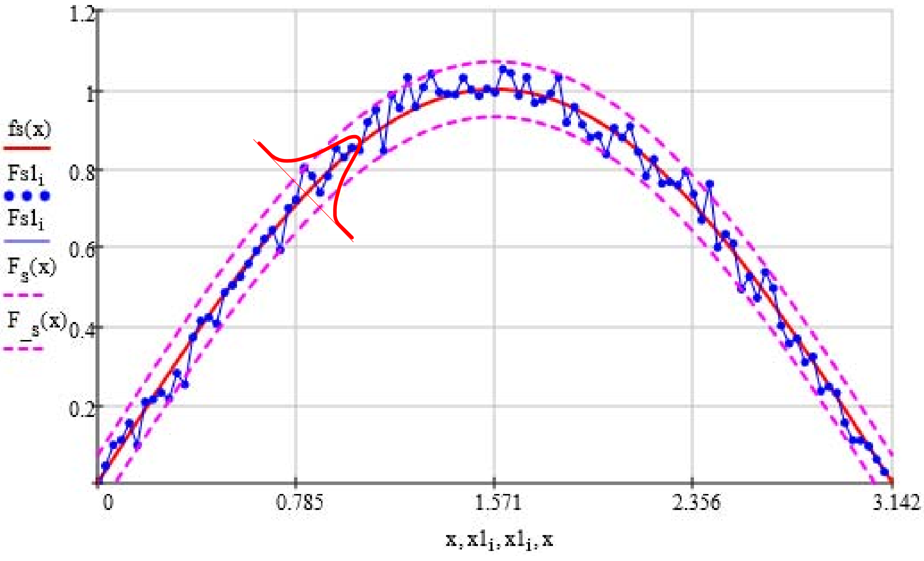
$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-3\sigma}^{3\sigma}e^{-\frac{(x-x_M)^2}{2\sigma^2}}=0,997,\quad F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-2\sigma}^{2\sigma}e^{-\frac{(x-x_M)^2}{2\sigma^2}}=0,95$$

Приведем пример двух моделей с гауссовым шумом.





Дисперсия 7%



Дисперсия 3.5%

Под моментами случайной величины понимается выражение вида

$$M_n = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} x^n e^{-\frac{(x-x_M)^2}{2\sigma^2}}, \quad M_0 = 1,\ M_1 = x_M,\ M_2 = \sigma^2$$

Обычно начиная со второго момента моменты центрируют

$$M_n = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} \left(x - x_M\right)^n e^{-\frac{(x-x_M)^2}{2\sigma^2}},$$

Как определяются моменты экспериментально? Моменты определяются по выборкам. По совокупности измеренных данных. Сначала нужно использовать регрессионный анализ и МНК для определения трендовой составляющей. По теореме Вейерштрасса искомое решение представляется в $\tilde{y}_k = a_0 + a_1 k + a_2 k^2 + a_3 k^3$

$$Q(a) = \sum_{i=0}^{N} \left(y_i - \tilde{y}(a_k, x_i)\right)^2 = \sum_{i=0}^{N} \left(y_i - a_3 x_i^3 - a_2 x_i^2 - a_2 x_i - a_0\right)^2 \to \min$$

Затем вычисляются частные производные по неизвестным коэффициентам

$$\frac{\partial Q}{\partial a_k} = 2\sum_{i=0}^{N} \left(y_i - \tilde{y}(a_k, x_i)\right) \frac{\partial \tilde{y}(a_k, x_i)}{\partial a_k} = 0$$

И решается система алгебраических уравнений относительно a_k. Всё это можно сделать с быстрее с помощью псевдо обратной матрицы Мура-Пенроуза $\mathbf{A}^+ = \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T$. Матрица **A** которая входит в псевдо обратную матрицу формируется по алгоритму.

$\mathbf{A}_{i,k} = \left(x_i\right)^k$, здесь *k*- степень полинома, *i*- число выборок. Запишем формулу коэффициентов аппроксимирующего полинома.

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \left(X^T X\right)^{-1} X^T y\,,$$

Здесь $y = \left(y_0, y_1, y_2.....y_N\right)^T$ совокупность табличных значений.

Затем составляется выражение для тренда

$$y(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

Величины, которые дают нам приближенное значение моментов через табличные данные называются оценками или выборочные моменты. Приведём формулы для оценок моментов. Математическое ожидание

$$\delta_k = y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3\right), \quad x_M = \frac{1}{N}\sum_{k=0}^{N-1}\delta_k$$

Оценка дисперсии определяется по формуле:

$$D = \frac{1}{N}\sum_{k=0}^{N-1}\left(\delta_k - x_M\right)^2$$

Рассмотрим **пример.** Получены табличные значения. По табличным значения необходимо определить закон распределения случайной величины.

$i := 0..N-1$

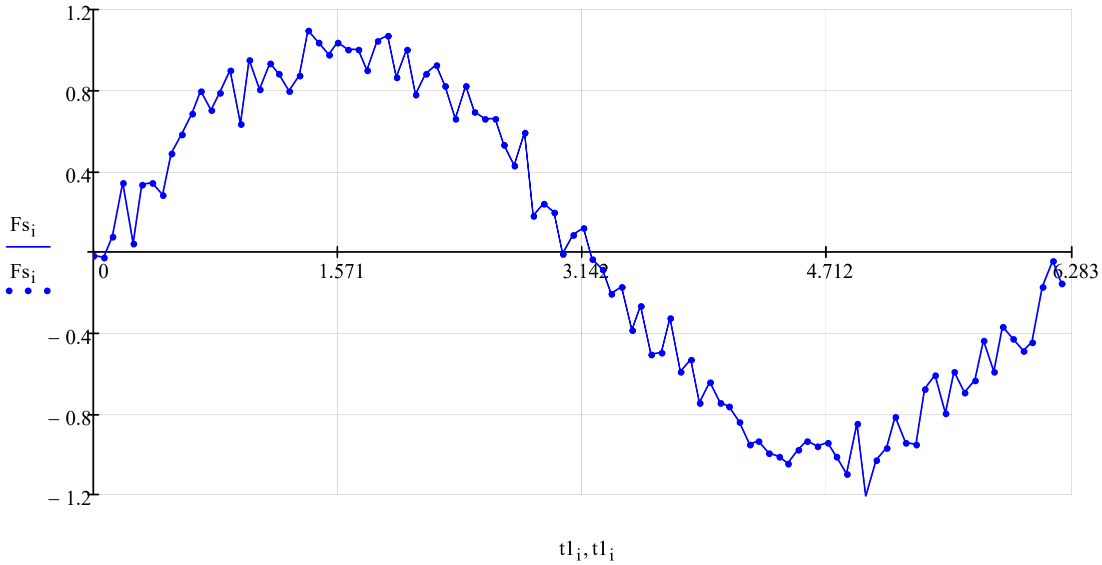
$t1^T =$

	0	1	2	3	4	5	6	7	8	9
0	0	0.063	0.126	0.188	0.251	0.314	0.377	0.44	0.503	...

$Fs^T =$

	0	1	2	3	4	5
0	-0.013	-0.028	0.08	0.346	0.041	...

По табличным данным строим графики.

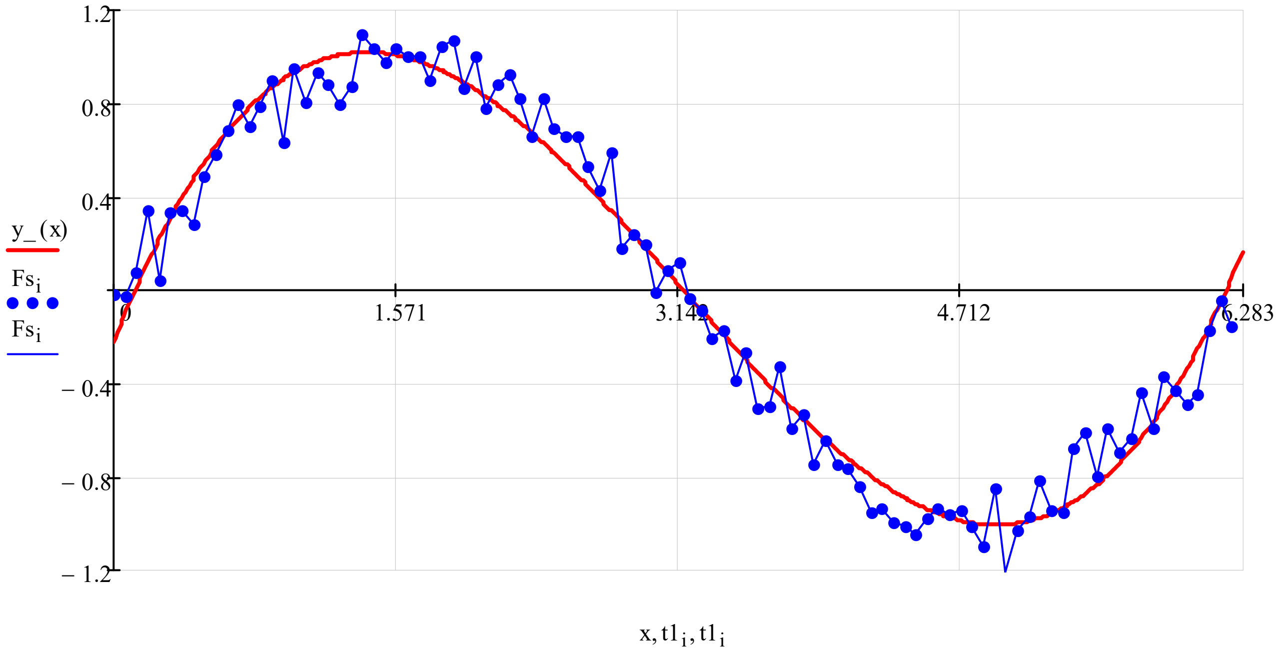


Определяем тренд, используя метод наименьших квадратов (МНК)

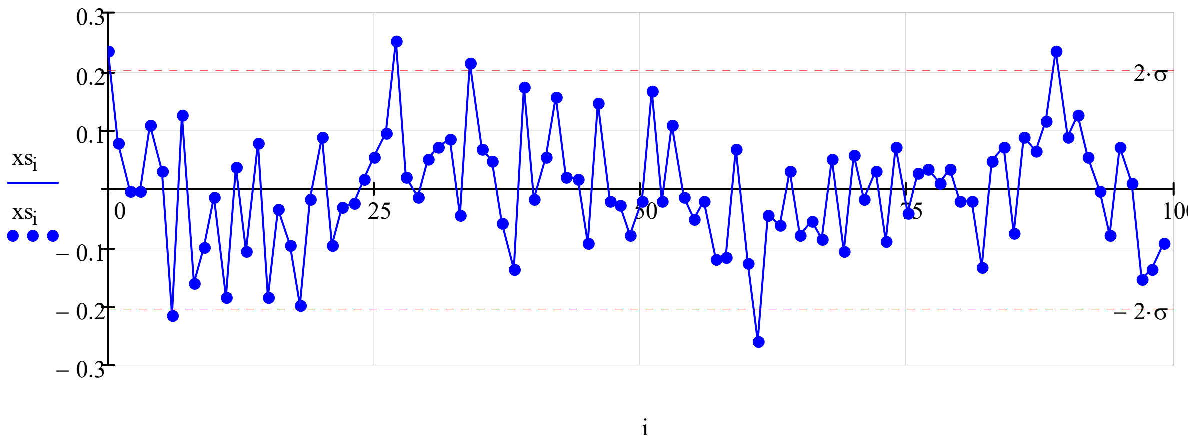
$$k := 0..3 \quad X_{i,k} := \left(x1_i\right)^k \quad \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} := \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Fs = \begin{pmatrix} -0.221 \\ 1.952 \\ -0.893 \\ 0.094 \end{pmatrix}$$

$$k := 0..3 \quad i := 0..N-1 \quad X_{i,k} := \left(t1_i\right)^k \quad \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} := \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Fs = \begin{pmatrix} -0.221 \\ 1.952 \\ -0.893 \\ 0.094 \end{pmatrix}$$

$$y\_(t) := a_0 + a_1 \cdot t + a_2 \cdot t^2 + a_3 \cdot t^3$$



Теперь можно определить случайную величину как отклонение от трендового значения $xs_i := \left(Fs_i - y\_\left(t1_i\right)\right)$ то есть величину ошибки



Находим оценки случайной величины, первый (выборочное среднее) и второй моменты.

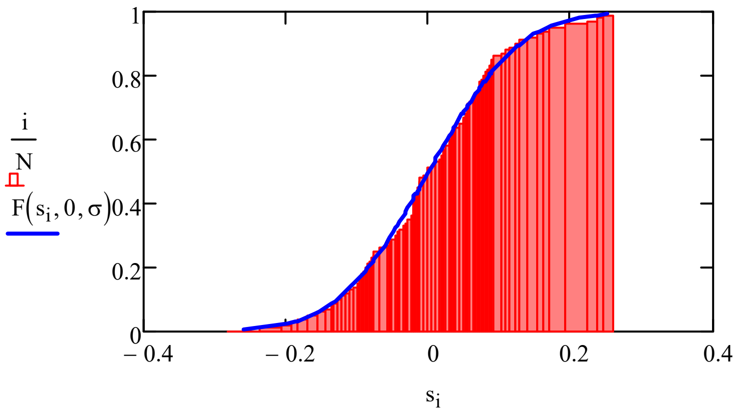
$$M_1 := \frac{1}{N}\cdot\sum_{i=0}^{N-1} xs_i = -0 \qquad M_2 := \frac{1}{N-1}\cdot\sum_{i=0}^{N-1}\left(xs_i - M_1\right)^2 = 0.0134$$

Определяем среднеквадратическое отклонение

$$D := M_2 \qquad \sigma := \sqrt{D} = 0.116$$

По графику видно, что выполняется закон 2-х и 3-х сигма.

$$s := \mathrm{sort}(xs) \qquad F(x, x_M, \sigma) := \int_{-100}^{x} f(x, \sigma, x_i)\,dx \qquad x := -2\cdot\sigma, -2\cdot\sigma + 0.025 .. 2.5\cdot\sigma$$



$N = 100 \qquad H := \mathrm{histogram}\left(\mathrm{trunc}\left(\sqrt{N}\right), s\right) \qquad H^{\langle 0\rangle} =$

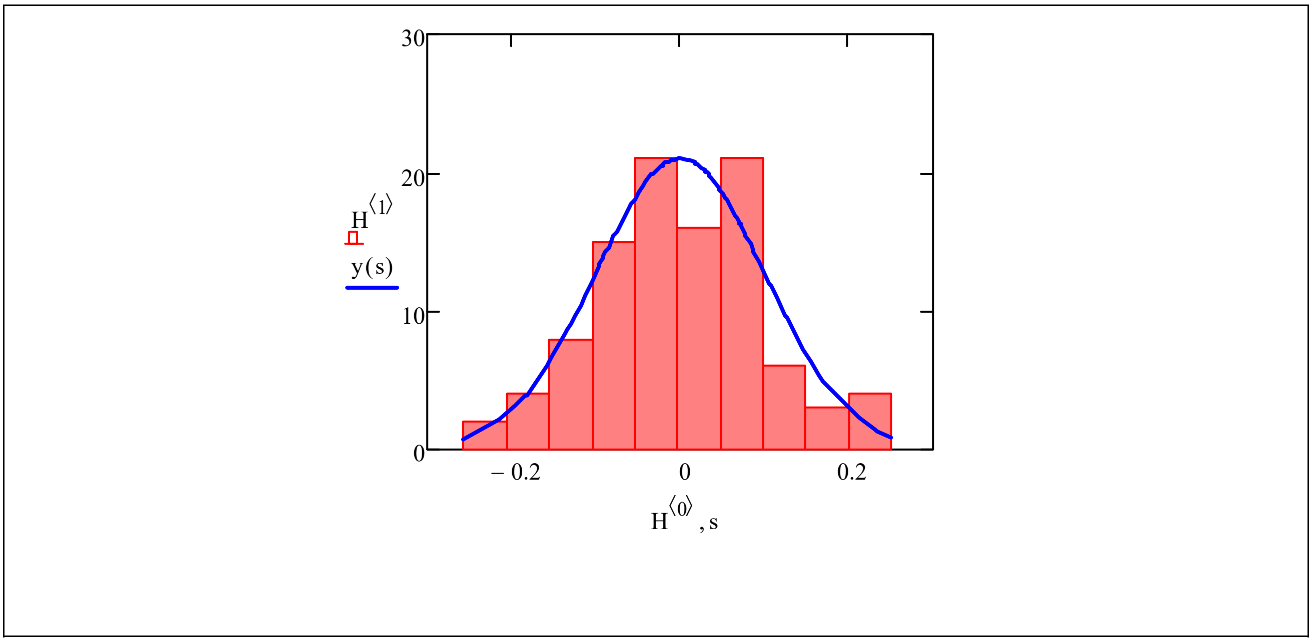
	0
0	-0.233
1	-0.182
2	-0.131
3	-0.08
4	-0.029
5	0.022
6	0.073
7	0.124
8	0.175
9	0.227

$H^{\langle 1\rangle} =$

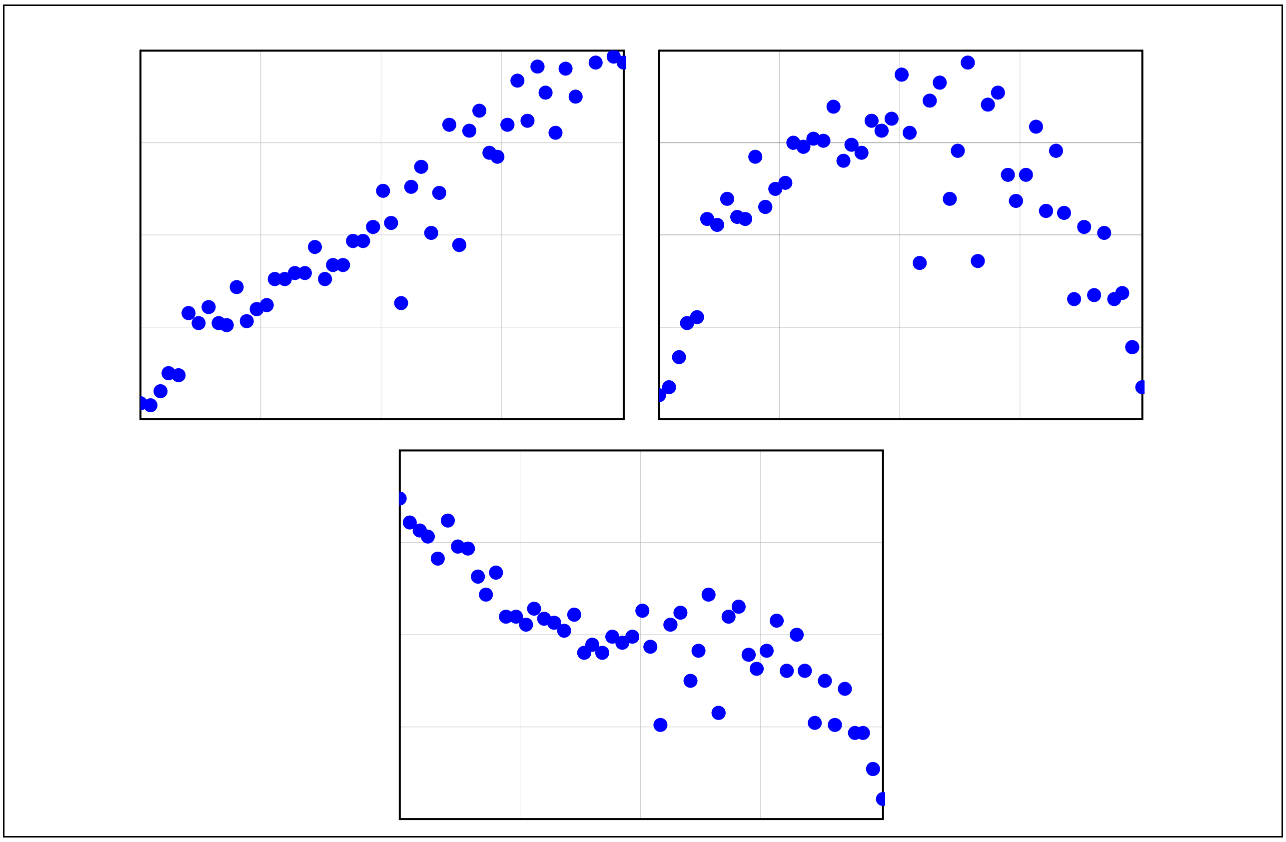
	0
0	2
1	4
2	8
3	15
4	21
5	16
6	21
7	6
8	3
9	4

Построение Гистограммы

$$y(x) := \mathrm{dnorm}(x, 0, \sigma)\cdot\sqrt{2\cdot\pi}\cdot\sigma\cdot\max\left(H^{\langle 1\rangle}\right)$$



Как проверить достоверность получаемых зависимостей



Расчет коэффициентов корреляции.

В качестве меры достоверности уравнения корреляционной зависимости используется процентное отношение средней квадратической ошибки уравнения *Se* к среднему уровню результативного признака *y* :