Подбор закона распределения

При обработке экспериментальных данных часто возникает необходимость проверить соответствие закона распределения одному из известных заонов. В частности при выборе методов анализа – параметрических/непараметрических необходимо убедиться в правильности гипотезы о соответствии/несоответствии нормальному заону. Для этой цели в *STATISTICA* предназначен модуль **Distribution Fitting** (подгонка распределения). Для изучения этого модуля воспользуемся файлом **Turtles.sta** из библиотеки **Examples**.

Отметим, что нулевая гипотеза H_0 формулируется так: закон распределения соответствует проверяемому распределению. Поэтому, если уровень значимости $p < 0,05$, то вероятность ошибиться, отклонив нулевую гипотезу мала, поэтому, ее отклоняем. Если $p > 0,05$ – гипотезу принимаем.

Чтобы запустить модуль **Distribution Fitting,** необходимо в главном меню **Statistics** выбрать одноименную команду. В открывшемся окне **Distribution Fiting** надо указать природу случайной величины, т.е. *Continuous Distribution* (непрерывная) или *Discret Distribution* (дискретная), а также предполагаемый закон распределения, которому случайная величина подчиняется.

Для непрерывных случайных величин предложено шесть законов распределения, а для дискретных – четыре (рис.2).

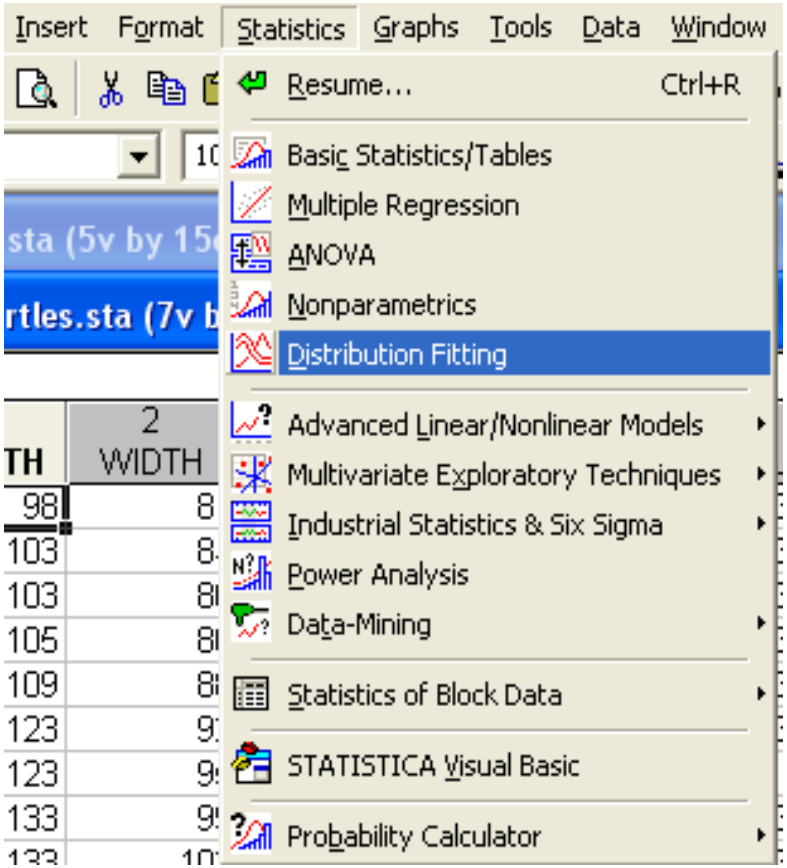


Рис.1

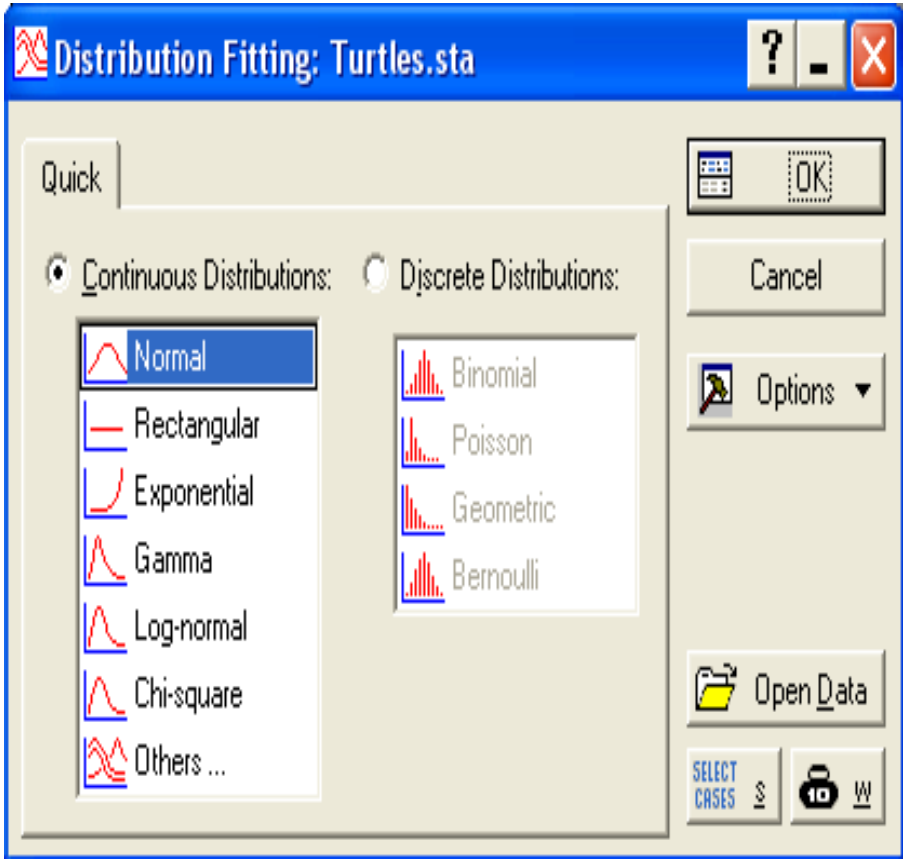


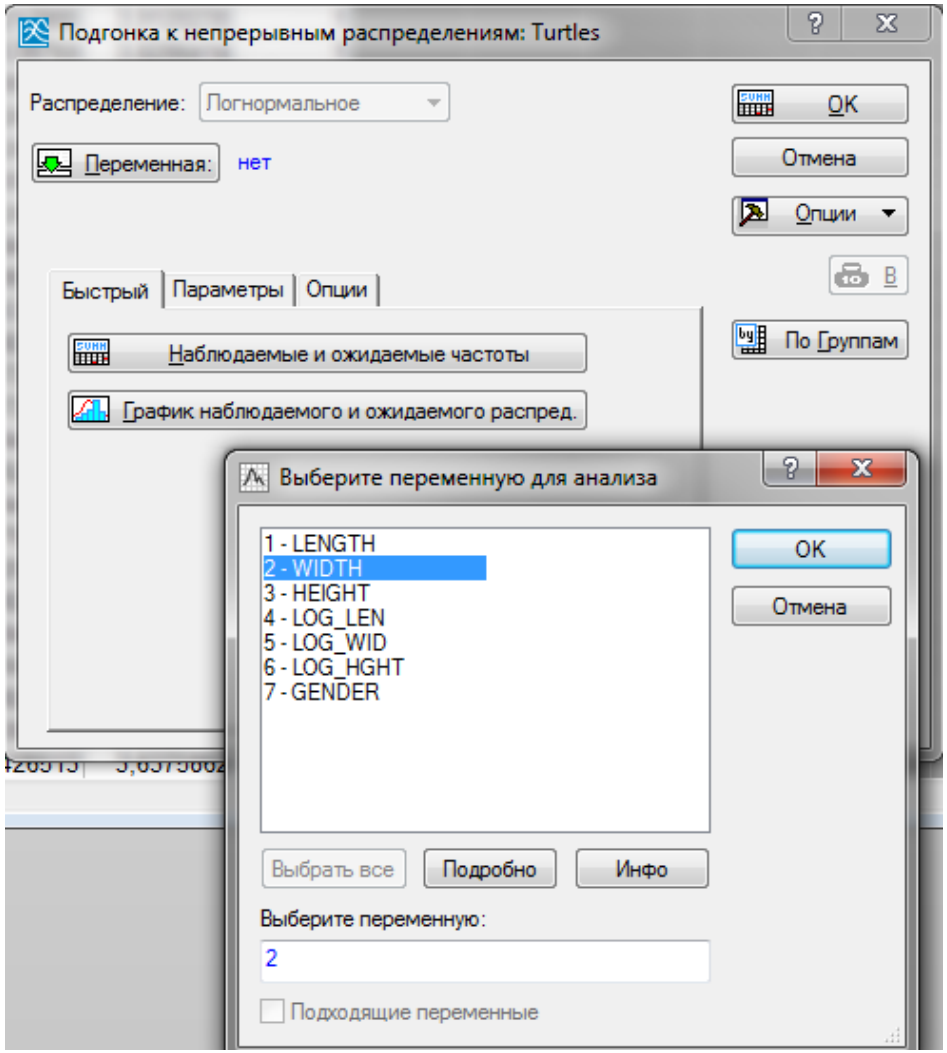
Рис.2

Для иллюстрации работы модуля воспользуемся таблицей Turtles.sta из папки Example

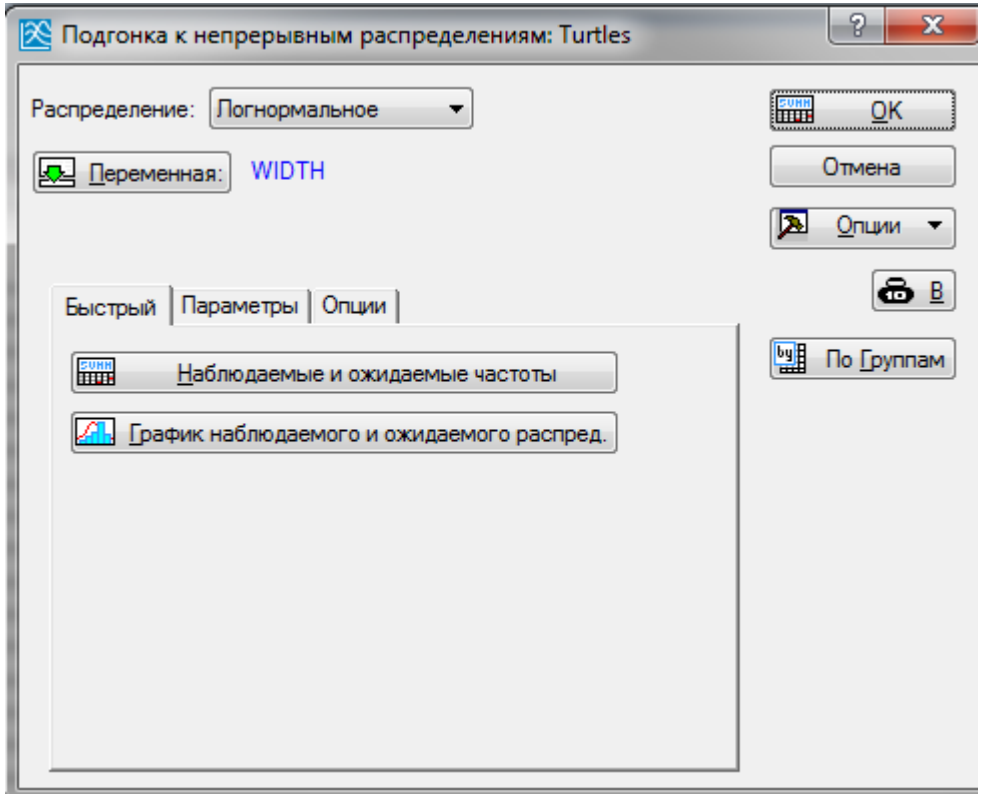
	LENGTH	WIDTH	HEIGHT	LOG_LEN	LOG_WID	LOG_HGHT	GENDER
1	98	81	38	4,5849675	4,3944492	3,6375862	1
2	103	84	38	4,6347290	4,4308168	3,6375862	1
3	103	86	42	4,6347290	4,4543473	3,7376696	1
4	105	86	42	4,6539604	4,4543473	3,7376696	1
5	109	88	44	4,6913479	4,4773368	3,7841896	1
6	123	92	50	4,8121844	4,5217886	3,9120230	1
7	123	95	46	4,8121844	4,5538769	3,8286414	1
8	133	99	51	4,8903491	4,5951199	3,9318256	1
9	133	102	51	4,8903491	4,6249728	3,9318256	1
10	133	102	51	4,8903491	4,6249728	3,9318256	1
11	134	100	48	4,8978398	4,6051702	3,8712010	1
12	136	102	49	4,9126549	4,6249728	3,8918203	1
13	138	98	51	4,9272537	4,5849675	3,9318256	1
14	138	99	51	4,9272537	4,5951199	3,9318256	1
15	141	105	53	4,9487599	4,6539604	3,9702919	1
16	147	108	57	4,9904326	4,6821312	4,0430513	1
17	149	107	55	5,0039463	4,6728288	4,0073332	1
18	153	107	56	5,0304379	4,6728288	4,0253517	1
19	155	115	63	5,0434251	4,7449321	4,1431347	1
20	155	117	60	5,0434251	4,7621739	4,0943446	1
21	158	115	62	5,0625950	4,7449321	4,1271344	1
22	159	118	63	5,0689042	4,7706846	4,1431347	1
23	162	124	61	5,0875963	4,8202816	4,1108739	1
24	177	132	67	5,1761497	4,8828019	4,2046926	1
25	93	74	37	4,5325995	4,3040651	3,6109179	2
26	94	78	35	4,5432948	4,3567088	3,5553481	2
27	96	80	35	4,5643482	4,3820266	3,5553481	2
28	101	84	39	4,6151205	4,4308168	3,6635616	2
29	102	85	38	4,6249728	4,4426513	3,6375862	2
30	103	81	37	4,6347290	4,3944492	3,6109179	2

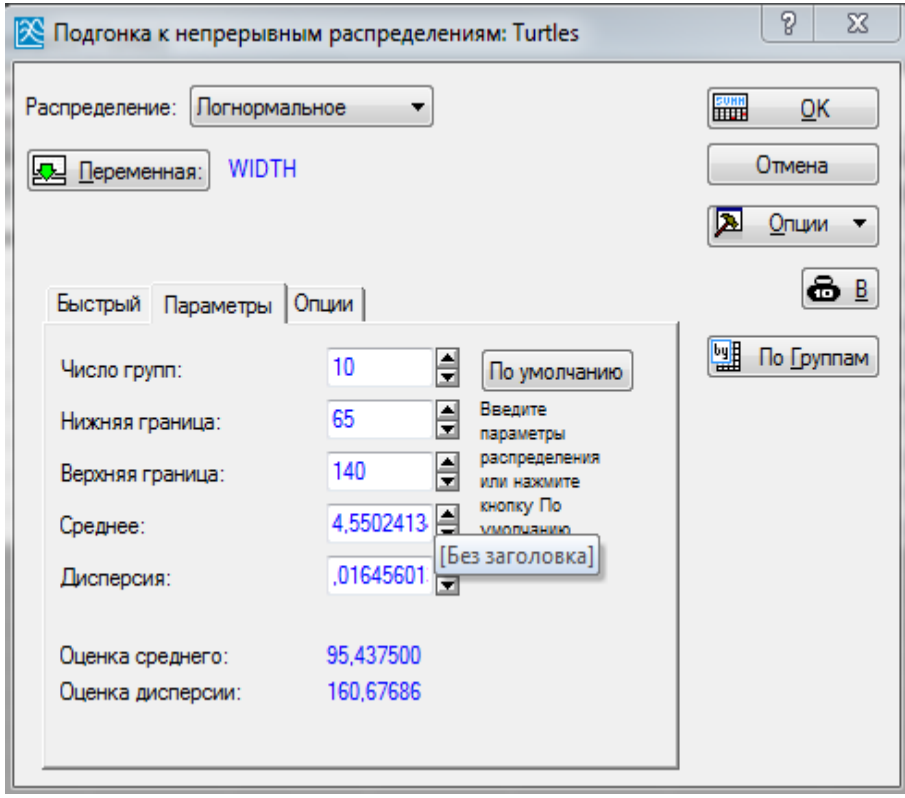
Выберите, например, *Lognormal* и нажмите **OK**

Щелкнув по кнопке Variable (Переменная) В открывшемся окне укажите переменную *WIDTH* (ширина), для которой будет производиться исследование. распределений)



Слева вверху становится активным выпадающее меню, и можно выбрать другой закон распределения. По умолчанию активной является вкладка **Quick.** На этой вкладке есть две кнопки: **Summary**: **Observed and expected distributions** (результат: наблюдаемые и ожидаемые частоты) и **Plot of Observed and expected distributions** (график наблюдаемых и ожидаемых. Перейдем на вкладку параметры и укажем число групп 10 вместо 15 по умолчанию





Если нажать на первую кнопку, то программа отобразит численные характеристики в виде таблицы (рис.3). Программа автоматически разбила диапазон **на 10 интервалов**. В первом столбце *Observed Frequency* (наблюдаемая частота) для каждого рассмотренного интервала указано количество значений, попавших в этот интервал. Во втором столбце *Cumulative Observed* (совокупный наблюдаемый) для каждого интервала приведено количество значений, попавших в этот и все предшествующие интервалы (накопленные частоты). В третьем и четвертом столбцах *Percent Observed* (процент наблюдаемый) и *Cumul. %* (суммарный процент) указаны те же величины, что и в предыдущих двух, но исчисленные в процентах.

	Variable: WIDTH, Distribution: Log-normal (Turtles.sta) Chi-Square = 2,33718, df = 2 (adjusted) , p = 0,31081								
Upper Boundary	Obs. Freq	Cumul Obser	Perc Obser	Cumul. % Obser	Expec Freq	Cumul Expec	Perc Expec	Cumul. % Expec	Obser Expec
<= 72,50000	0	0	0,00	0,00	0,90	0,90	1,88	1,88	-0,90
80,00000	3	3	6,25	6,25	3,65	4,55	7,61	9,49	-0,65
87,50000	11	14	22,92	29,17	8,41	12,96	17,51	27,00	2,59
95,00000	15	29	31,25	60,42	11,58	24,54	24,13	51,13	3,42
102,50000	8	37	16,67	77,08	10,62	35,16	22,13	73,26	-2,62
110,00000	5	42	10,42	87,50	7,04	42,20	14,66	87,92	-2,04
117,50000	3	45	6,25	93,75	3,59	45,79	7,48	95,40	-0,59
125,00000	2	47	4,17	97,92	1,48	47,28	3,09	98,49	0,52
132,50000	1	48	2,08	100,00	0,51	47,79	1,07	99,56	0,49
< Infinity	0	48	0,00	100,00	0,21	48,00	0,44	100,00	-0,21

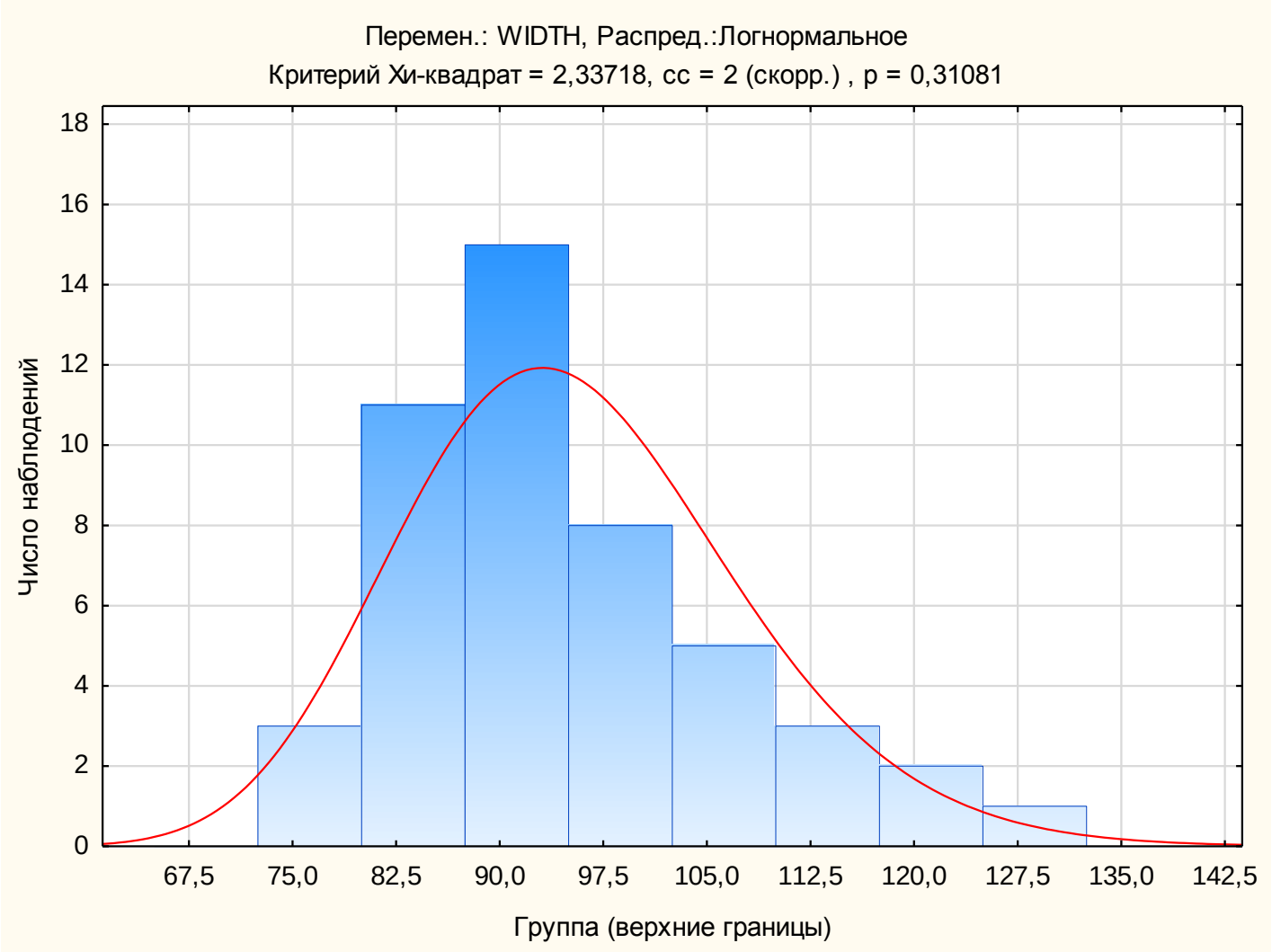
В пятом столбце *Frequency Expected* (ожидаемая частота) даны теоретические частоты, соответствующие логнормальному распределению. **В последнем столбце указана разность между исходной и ожидаемой накопленной частотой**. При нажатии на вторую кнопку будет построена кривая теоретического закона распределения и гистограмма эмпирического, построенного по имеющимся данным (рис.4). Над гистограммой выведен заголовок, в котором указана анализируемая переменная, предполагаемый закон распределения, а также три числовых параметра, которые рассмотрим подробнее.

Первый параметр – это значение критерия $\chi 2$. Чем меньше это значение, тем больше вероятность того, что проверяемая случайная величина имеет предполагаемый закон распределения.

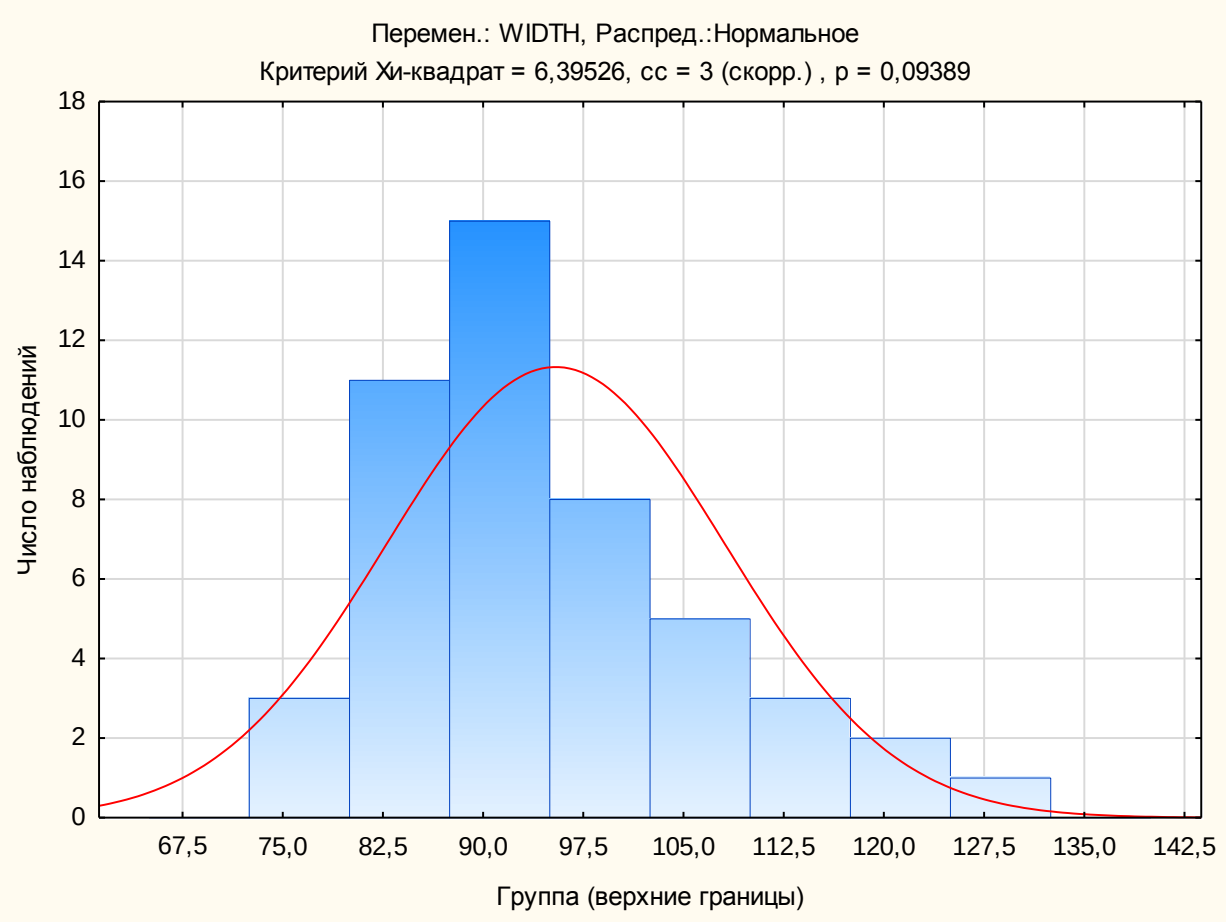
Второй параметр *df (cc)* – число степеней свободы. Определяется как $df = n\text{-}l\text{-}1$, где ***n* – число интервалов, в которых ожидаемые частоты более 5. Если есть интервалы с меньшим количеством ожидаемых частот, то программа объединит их**; l – число оцениваемых параметров распределения (для логнормального рас пределения $l = 2$). В нашем случае после объединения из 10 интервалов останется 5. Тогда $df = 5\text{-}2\text{-}1=2$

Третий параметр *p-уровень* значимости критерия, который определяет вероятность ошибки при отклонении гипотезы о нормальности. Так как вероятность ошибки достаточно велика, примерно 0,31 (что значительно больше 0,05), гипотезу о соответствии закона распределения логнормальному принимаем.

Если щелкнем по кнопке график наблюдаемых и ожидаемых значений, то появится графическое изображение закона распределения, на котором видно (p>0,05), что кривая распределения обладает незначительной асимметрией



При выборе нормального распределения будет построен график уже с симметричной кривой распределения. Как видно (p>0,05), также подтверждается гипотеза о соответствии закона распределения нормальному! Но при этом p значительно меньше, чем при логнормальном распределении, поэтому правильно считать верной гипотезу о соответствия распределения логнормальному закону



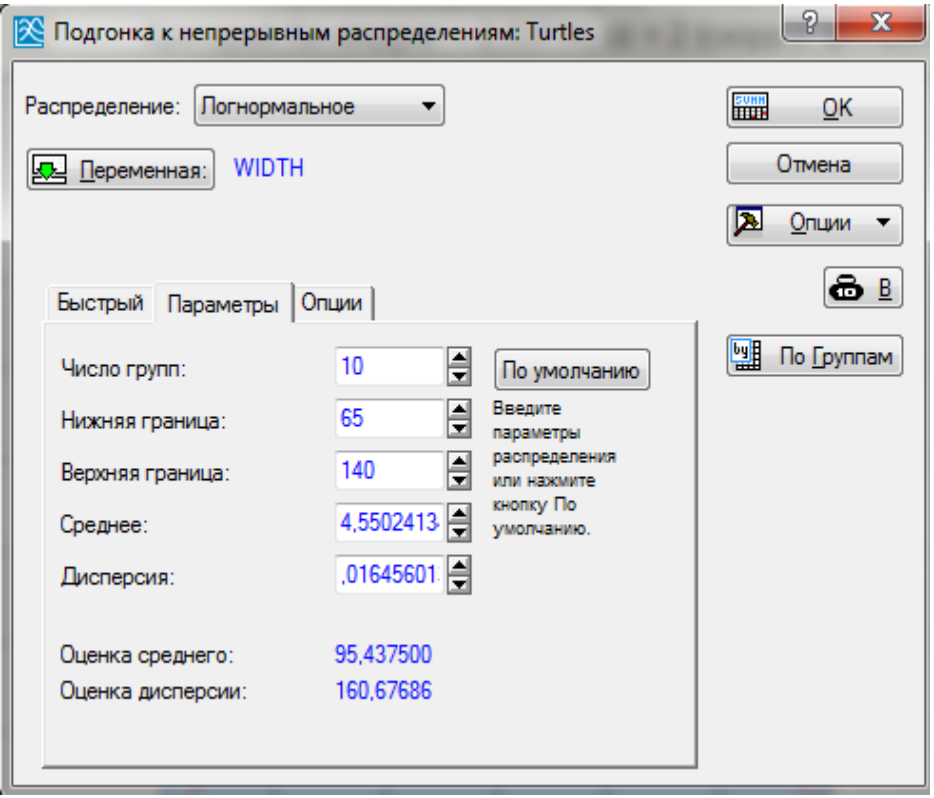
Кратко опишем другие вкладки рабочего окна рассматриваемого модуля.

На вкладке **Parameters** приведены значения параметров предполагаемого закона распределения. Кнопка **Set To Default** – установить значения по умолчанию. Среди приведенных здесь параметров три являются общими для всех распределений, а остальные зависят от выбора распределений.

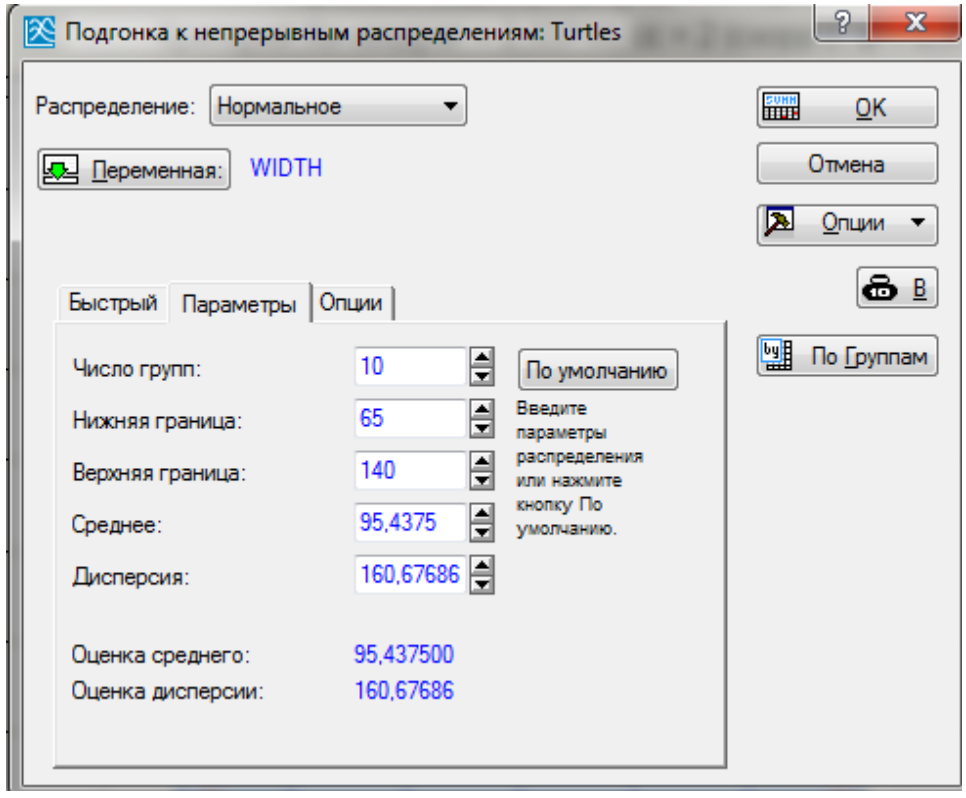
Number of categories (количество категорий) – количество интервалов, на которое будет разбита выборка.

Lower Limit, Uper Limit (нижний и верхний пределы). По умолчанию берутся, минимальное и максимальное значения выборки соответственно, однако изменив эти параметры, можно исключить из рассмотрения все значения, не попадающие в интересующий нас интервал.

Mean and Variance (среднее и дисперсия). Эти параметры определяются программой автоматически, но их можно переопределить и вручную, если, например, требуется не определить закон распределения, а проверить, насколько распределение случайной величины отличается от закона распределения с заданными параметрами.



Для логнормального распределения



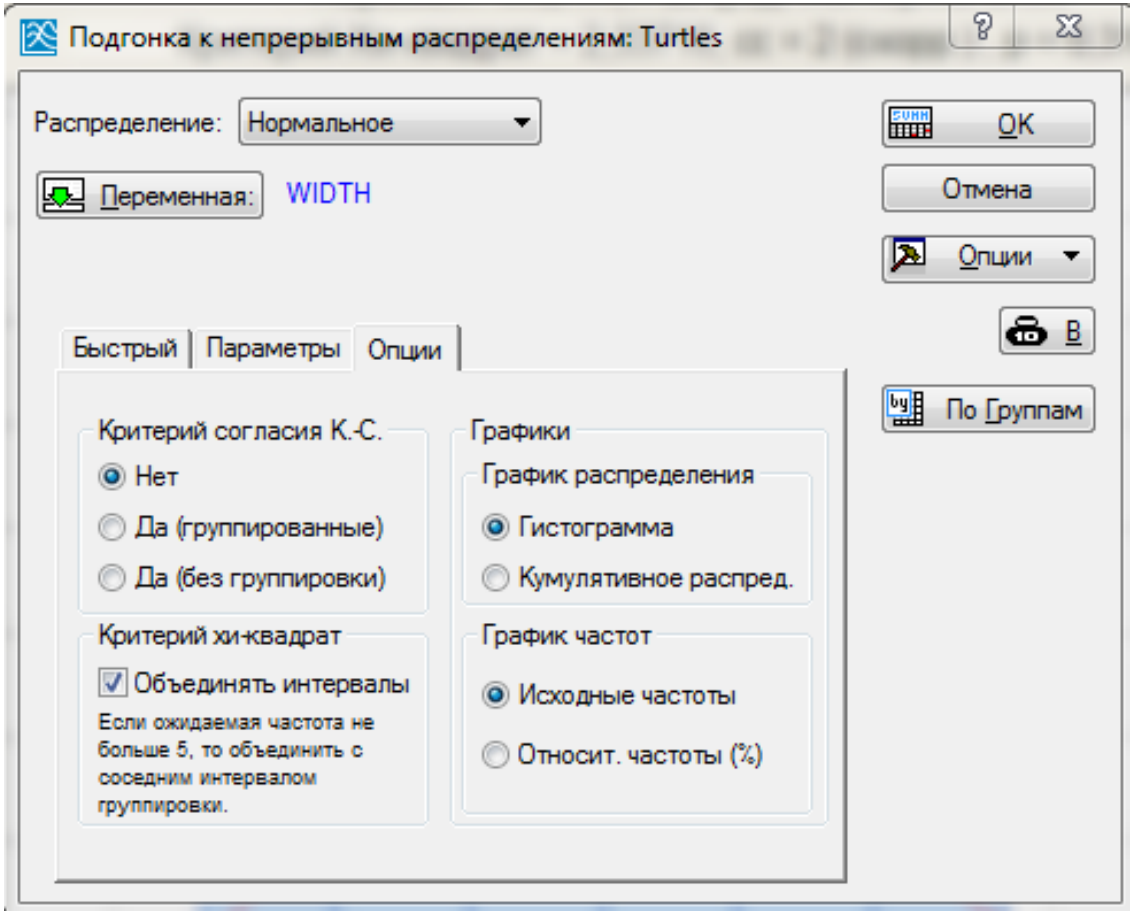
Для нормального распределения

На вкладке **Options** отображены настройки:

1. Kolmogorov – Smirnov test – критерий Колмогорова – Смирнова проверки гипотезы о соответствии выборочных данных тому или иному закону распределения. Статистика Колмогорова вычисляется на основании максимальной абсолютной разности между гипотетической функцией распределения и эмпирической функцией распределения.

Можно выбрать три опции: No (тест не вычисляется,) вычисляется по группированным (интервальным) данным, вычисляется по не группированным данным.

2. Chi – Square test. Критерий $\chi 2$ (Пирсона) проверки гипотезы о соответствии выборочных данных тому или иному закону распределения. Если в интервал попало менее 5 значений, при установке флажка на *Combine Categories* он объединяется с соседним и т.д., пока количество значений в интервалах будет не менее 5. Для нового разбиения вычисляется значение критерия $\chi 2$. В противном случае интервалы не объединяются.



Рамка *Graph Plot Distributions*. Если установить флажок на *Frequency distribution*, то программа построит график плотности распределения. Если флажок установить на *Cumulative distributions*, будет построен график функции распределения.

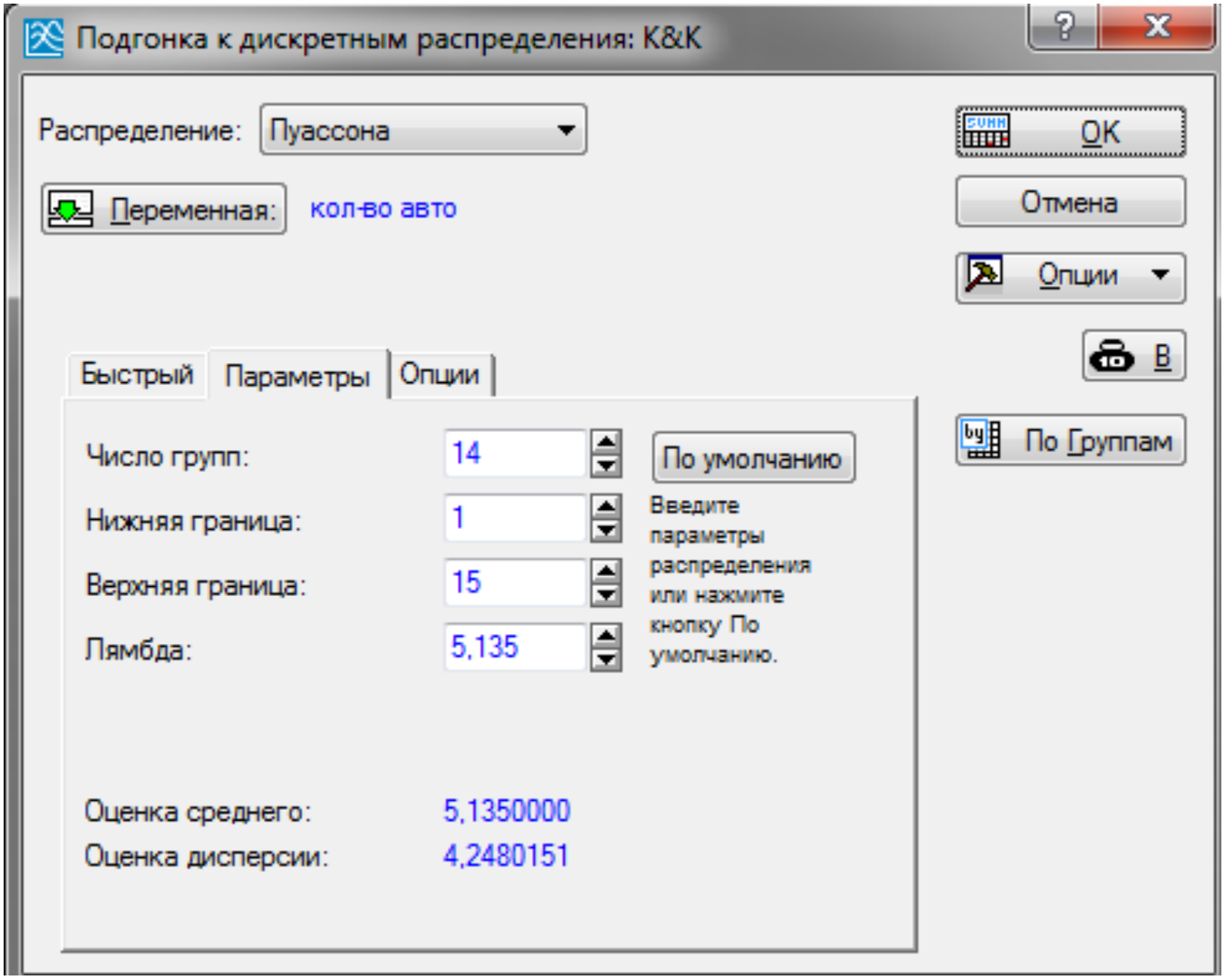
Рамка *Plot row frequencies or %*. Если установить флажок на *Raw frequencies*, на вертикальной оси графика будут отложены значения абсолютных частот, в противном случае – относительных (%).

Рассмотрим пример приближения эмпирического дискретного распределения пуассоновским. Ежедневно десятки автомобилей фирмы К&К в течение дня развозят продукцию компании в различные торговые организации. Для уменьшения времени простоя автомобилей в очереди при загрузке было решено изучить закон распределения количества автомобилей, подъезжающих к складским помещениям в течение часа, что позволит определить оптимальное количество кладовщиков, грузчиков, погрузочных площадок для организации эффективной работы складов. С этой целью 200 раз было подсчитано количество автомобилей, подъехавших в течение часа к складским помещениям. Данные файла К&К, которые представляют собой один столбец с 200 элементами, приведены на рис.7 в виде прямоугольной таблицы.

	Количество автомобилей, подъехавших в течении часа									
	1	2	3	4	5	6	7	8	9	10
1	9	6	7	5	8	3	4	6	3	3
2	8	5	5	3	5	5	5	7	6	6
3	6	3	7	5	6	3	4	5	6	6
4	4	8	5	3	7	1	5	6	3	3
5	3	9	5	5	5	5	4	6	5	5
6	5	4	5	8	5	7	7	9	4	4
7	5	2	6	8	5	6	5	5	6	6
8	5	7	5	5	13	3	7	3	3	3
9	8	3	3	5	8	2	4	7	3	3
10	6	7	2	7	5	3	6	4	2	2
11	4	8	6	10	4	4	3	3	6	6
12	6	4	7	1	4	3	2	4	6	6
13	4	3	3	3	4	3	3	4	7	7
14	3	5	3	4	3	4	5	3	9	9
15	6	4	5	5	6	5	6	2	9	9
16	7	4	6	5	2	1	6	5	4	4
17	9	8	7	8	6	8	7	5	6	6
18	4	2	3	3	4	8	3	7	6	6
19	10	7	2	5	5	6	7	4	3	3
20	4	7	3	4	10	2	5	6	9	9

Рис.7

В главном меню **Statistics** выберите команду **Distribution Fitting.** В открывшемся окне надо указать вид случайной величины, а именно – *Discret Distribution* (дискретная). В списке дискретных распределений укажите предполагаемый закон распределения, например *Poisson.* Щелкните по **OK**. В открывшемся окне на вкладке **Quick** нажмите кнопку **Plot of Observed and expected distributions.** Программа построит гистограмму – график эмпирической плотности распределения – и обозначит красной линией кривую предполагаемого теоретического распределения (рис.8). Уровень значимости критерия p = 0,149 принимает большее, чем 0,05 значение, чтобы можно было отвергнуть гипотезу о соответствии закона распределения пуассоновскому. В информационном поле указано значение параметра λ = 5,135, равное среднему числу автомобилей, находящихся в течение часа в очереди на загрузку.


Подгонка к дискретным распределения: K&K
Распределение:
Пуассона
OK
Переменная:
кол-во авто
Отмена
Опции
В
Быстрый
Параметры
Опции
Число групп:
14
По умолчанию
По Группам
Нижняя граница:
1
Введите параметры распределения или нажмите кнопку По умолчанию.
Верхняя граница:
15
Лямбда:
5,135
Оценка среднего:
5,1350000
Оценка дисперсии:
4,2480151

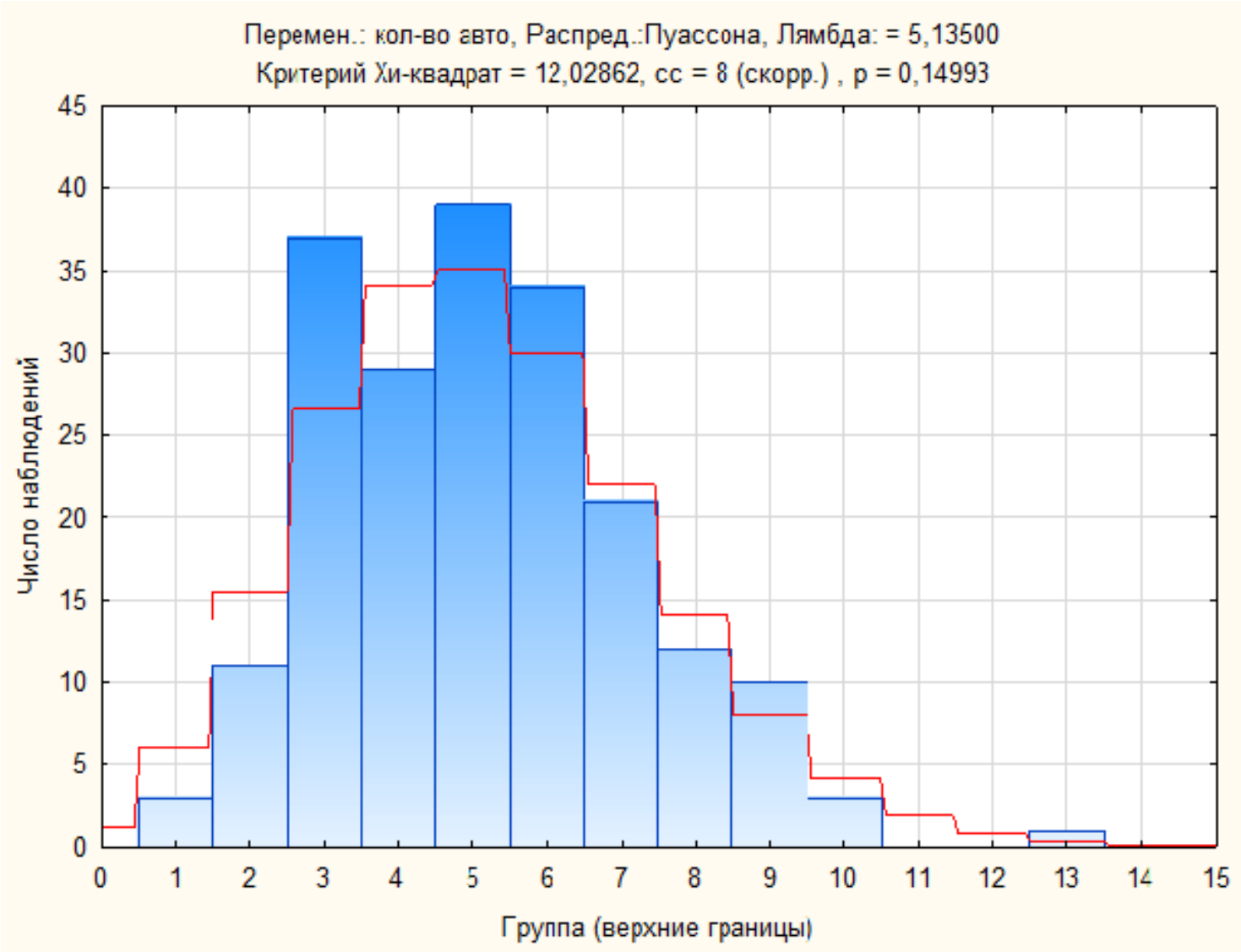

Перемен.: кол-во авто, Распред.:Пуассона, Лямбда: = 5,13500
Критерий Хи-квадрат = 12,02862, сс = 8 (скорр.) , p = 0,14993
Число наблюдений
Группа (верхние границы)
0
5
10
15
20
25
30
35
40
45
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Рис.8

- Так как модели теории массового обслуживания предполагают соответствие закона распределения входного потока заявок на обслуживание распределению Пуассона, проведенное исследование позволит использовать методы теории массового обслуживания для минимизации очереди автомобилей при загрузке товара.
- В рассмотренном примере программа анализировала закон распределения для данных, представленных в виде не сгруппированного ряда – перечислены значения случайной величины в порядке их появления в выборке. В то же время программа может анализировать данные, прошедшие обработку и записанные в виде сгруппированного ряда – перечислены ранжированные значения случайной величины и соответствующие им частоты в выборке.

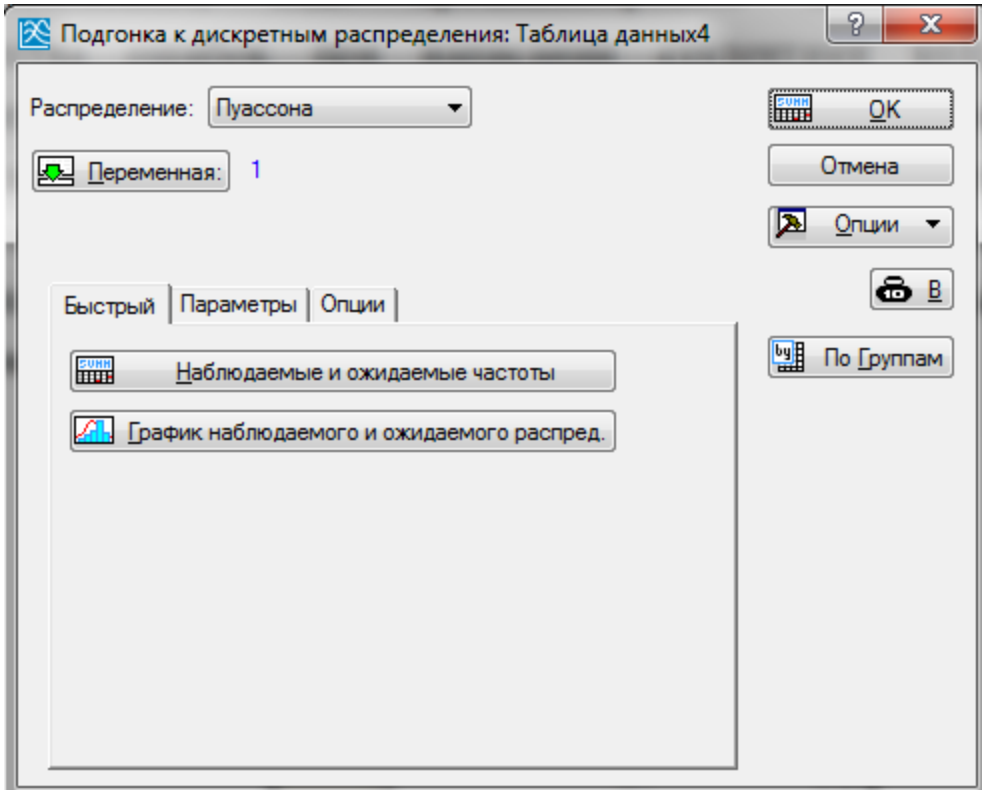
Предположим, надо проанализировать закон распределения количества ошибок при написании 100 студентами коллоквиума по математическому анализу. Сгруппированный ряд количества ошибок представлен в виде таблицы на рис.9. Как и в предыдущем случае в главном меню **Statistics** выберите команду **Distribution Fitting.** В открывшемся окне укажите вид случайной величины – **Discreet Distribution**.

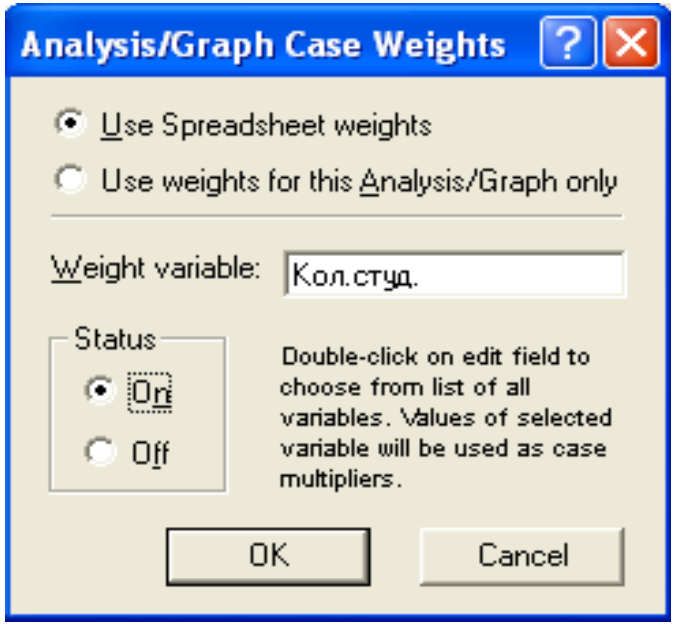
	1 Кол.ошиб.	2 Кол.студ
1	0	2
2	1	10
3	2	27
4	3	32
5	4	23
6	5	6

Рис.9

Воспользуемся кнопкой Variable, в открывшемся окне укажем имя *кол-во ошиб*.

В списке дискретных распределений укажем предполагаемый закон распределения, например, распределение Пуассона (*Poisson*). **Далее в правой части окна щелкнем на изображение гирьки** и в открывшемся окне в поле **Weight variable** наберем имя переменной *Кол.студ.* и произведем установки опций в соответствии с рис.10.





Щелкните по **OK** и, вернувшись в окно **Fitting Discrete Distributions,** нажмите кнопку **Plot of Observed and expected distributions.** Программа построит гистограмму (рис.11) эмпирического распределения с нанесенной на нее ломаной линией пуассоновского распределения.

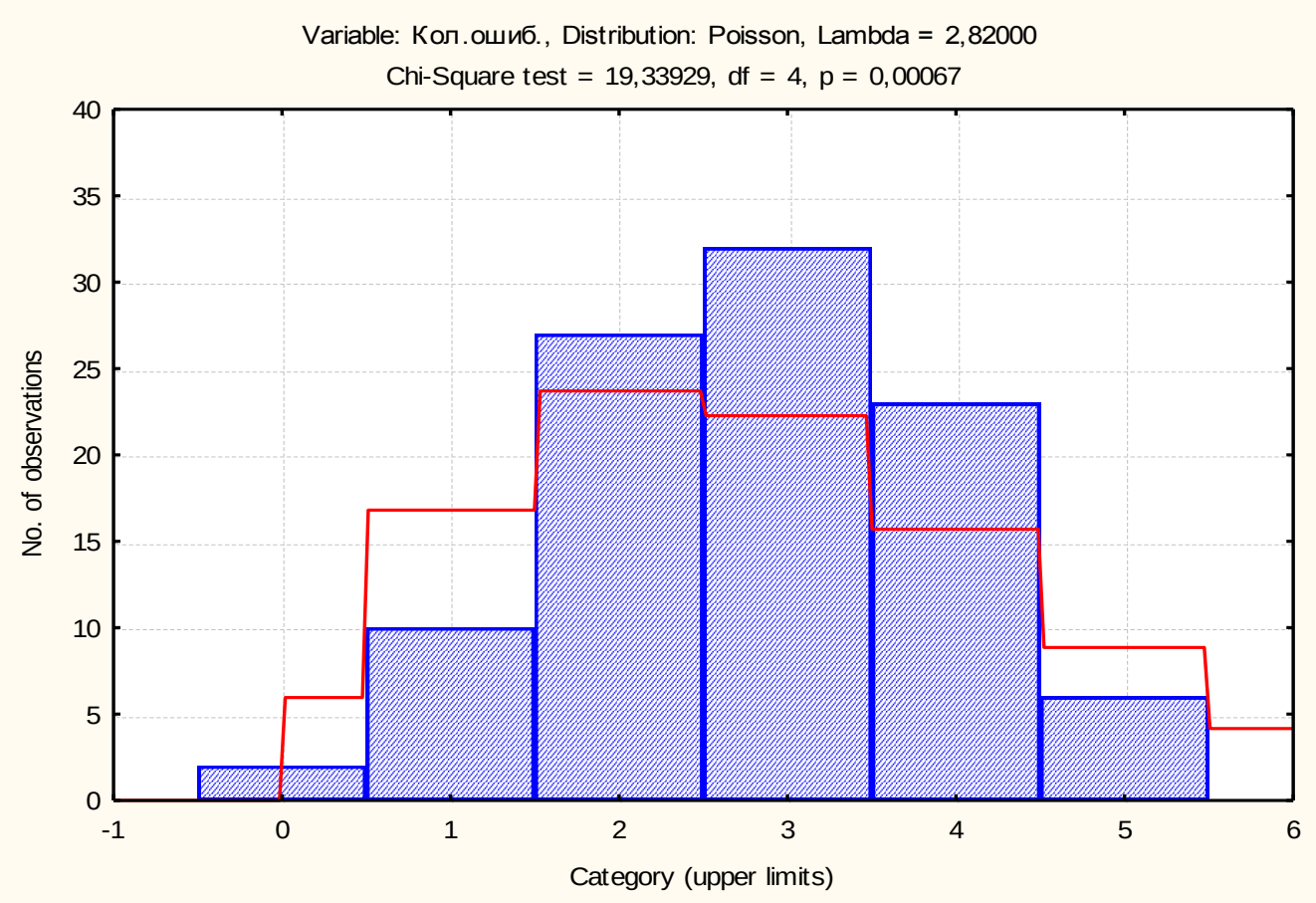


Рис.11

Из графиков, значений $\chi 2 = 19{,}33$ и уровня значимости вероятности $p=0{,}00007$ следует вывод о несоответствии закона распределения количества ошибок распределению Пуассона.

Повторим всю изложенную процедуру, указав в списке дискретных распределений – *Binomial*. Программа построит новый график (рис.12), из которого видно соответствие гистограмм; значение $\chi 2 = 0{,}205$; $p = 0{,}978$.

Малое значение $\chi 2$ и большое значение p говорят о большой вероятности ошибки, если отвергнуть гипотезу о соответствии закона распределения количества ошибок биномиальному закону. Поэтому гипотезу принимаем.

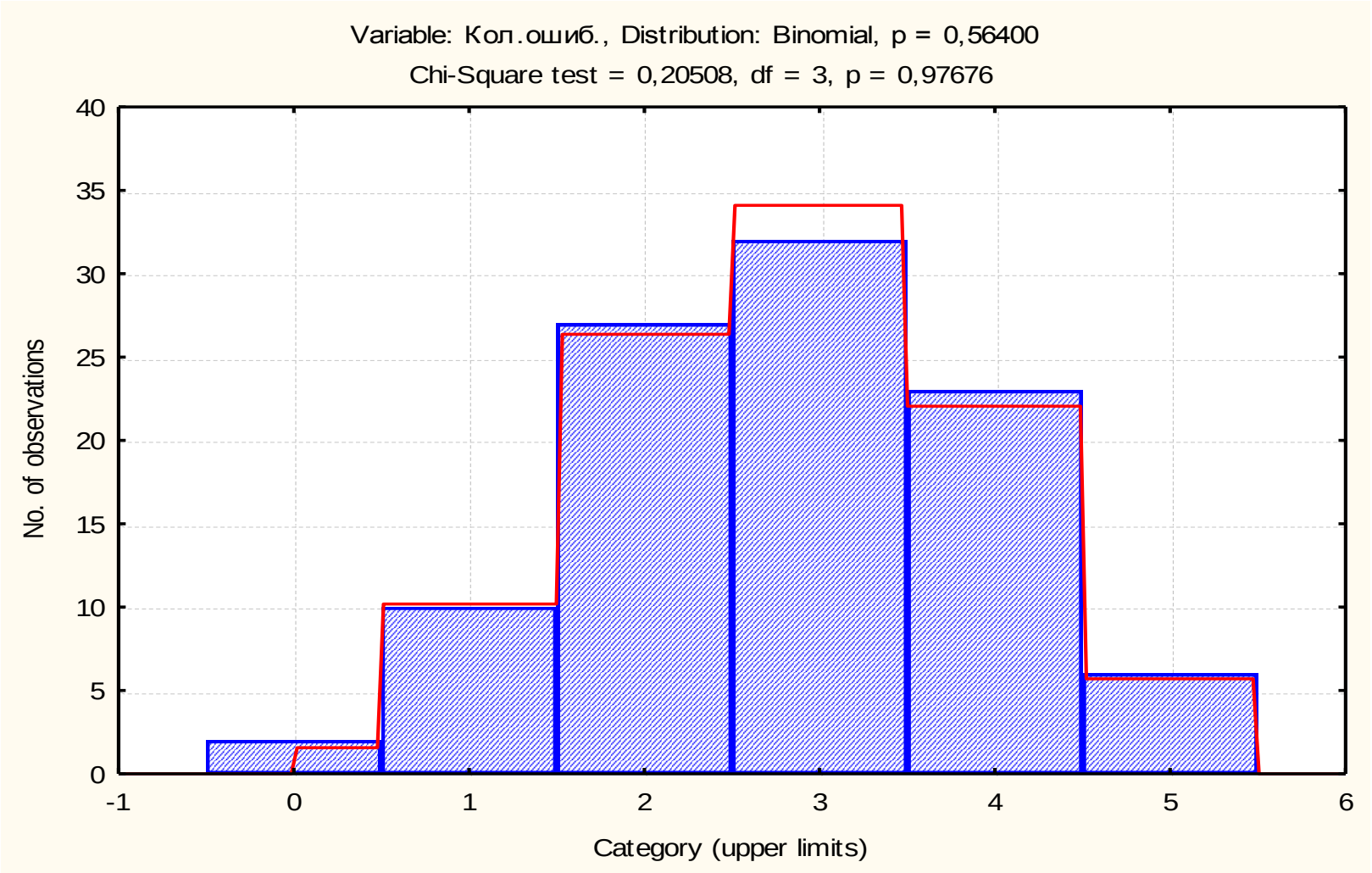

Variable: Кол.ошиб., Distribution: Binomial, p = 0,56400
Chi-Square test = 0,20508, df = 3, p = 0,97676
No. of observations
Category (upper limits)
40
35
30
25
20
15
10
5
0
-1
0
1
2
3
4
5
6

Рис.12

Генерация случайных чисел

В программе *STATISTICA* имеется возможность генерировать случайные числа, подчиняющиеся равномерному, нормальному и Пуассоновскому законам распределения. Известно, что с увеличением объема выборки возрастает соответствие эмпирического закона распределения теоретическому. Так, например, если количество генерируемых чисел более 1000, то отклонение эмпирического закона от теоретического практически незаметно. Если генерируется примерно 500 чисел, то видны отклонения от закона распределения. При количестве случайных величин менее 100 отклонения значительные и охарактеризовать выборку каким-либо законом распределения весьма затруднительно.

Для генерации случайных чисел надо дважды щелкнуть в таблице данных (в которой предполагается записать сгенерированные числа) на имени переменной. В окне спецификации переменной нажмите кнопку **Functions.** В открывшемся окне надо выделить *All Functions* и выбрать нужную функцию.

RND(X) (генерация равномерно распределенных чисел, рис.13). Эта функция имеет только один параметр – *X*, который задает правую границу интервала, содержащего случайные числа. При этом 0 является левой границей.

Аналогичные действия выполняет и функция *Uniform(X).* Чтобы вписать общий вид функции *RND(X)* в окно спецификации переменной, достаточно дважды щелкнуть на имени функции в окне **Function Browser.** После указания числового значения параметра *X* надо нажать **OK.** Для изменения значения математического ожидания на величину **g** надо заменить общий вид функции на *RND(X) + g.*

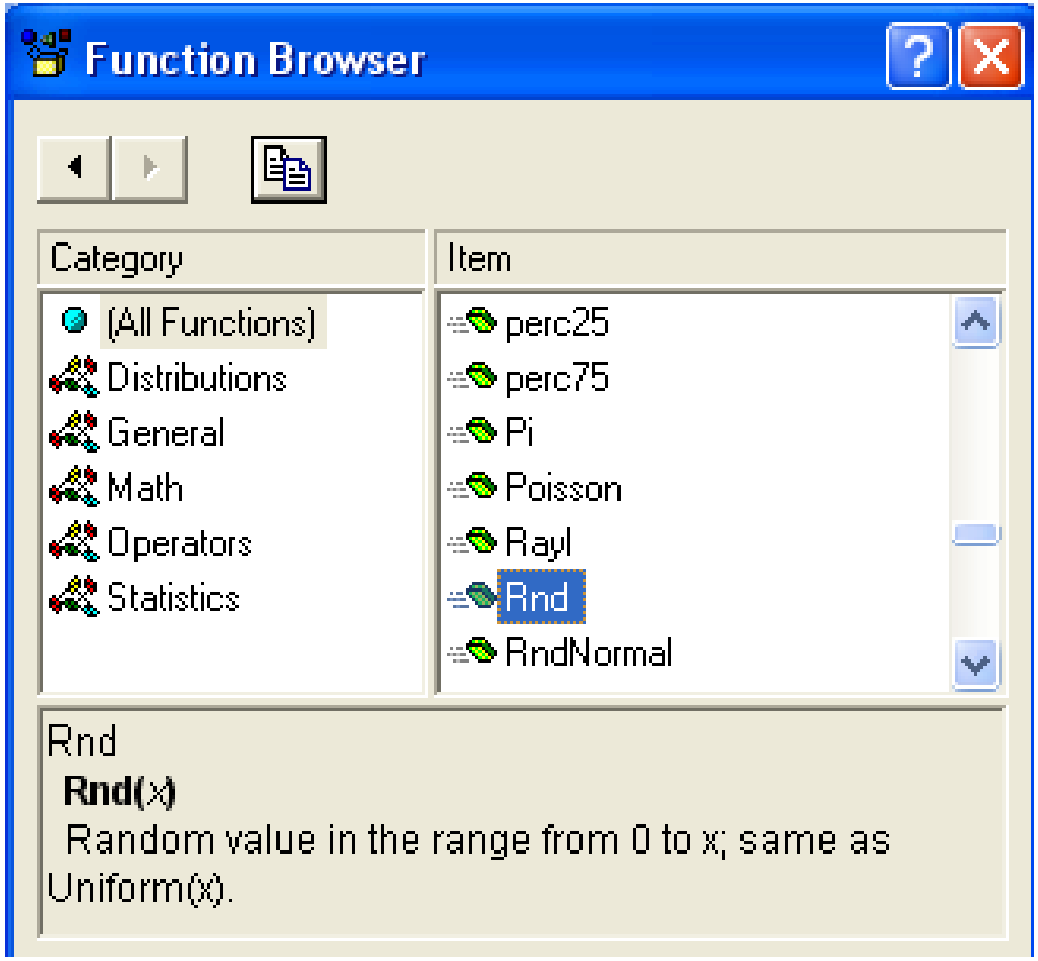

Function Browser
Category
Item
(All Functions)
Distributions
General
Math
Operators
Statistics
perc25
perc75
Pi
Poisson
Rayl
Rnd
RndNormal
Rnd
Rnd(x)
Random value in the range from 0 to x; same as Uniform(x).

Рис.13

RNDNormal (генерация нормально распределенных чисел, рис.14). Эта функция имеет один параметр – *X,* соответствующий стандартному отклонению случайной величины с математическим ожиданием 0. Запись *RNDNormal(X) + g* означает генерацию чисел с математическим ожиданием *g*.

RndPoisson (генерация чисел, соответствующих распределению Пуассона, рис.15). Функция имеет один параметр – *X,* соответствующий среднему значению. При необходимости генерирования случайных чисел других законов распределения надо воспользоваться известными в теории вероятностей и математической статистике соотношениями.

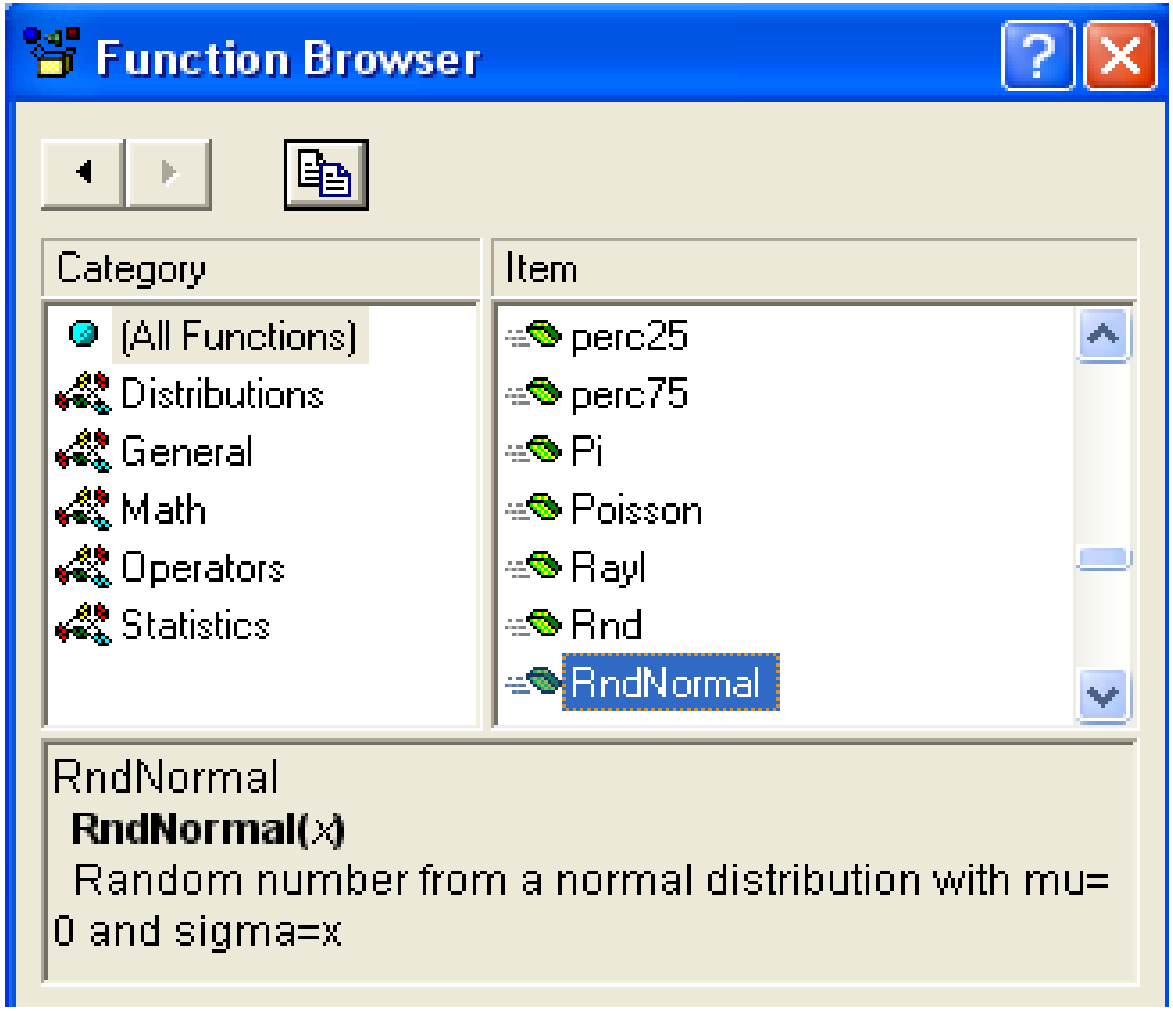

Function Browser
Category
Item
(All Functions)
Distributions
General
Math
Operators
Statistics
perc25
perc75
Pi
Poisson
Rayl
Rnd
RndNormal
RndNormal
RndNormal(x)
Random number from a normal distribution with mu=
0 and sigma=x

Рис.14

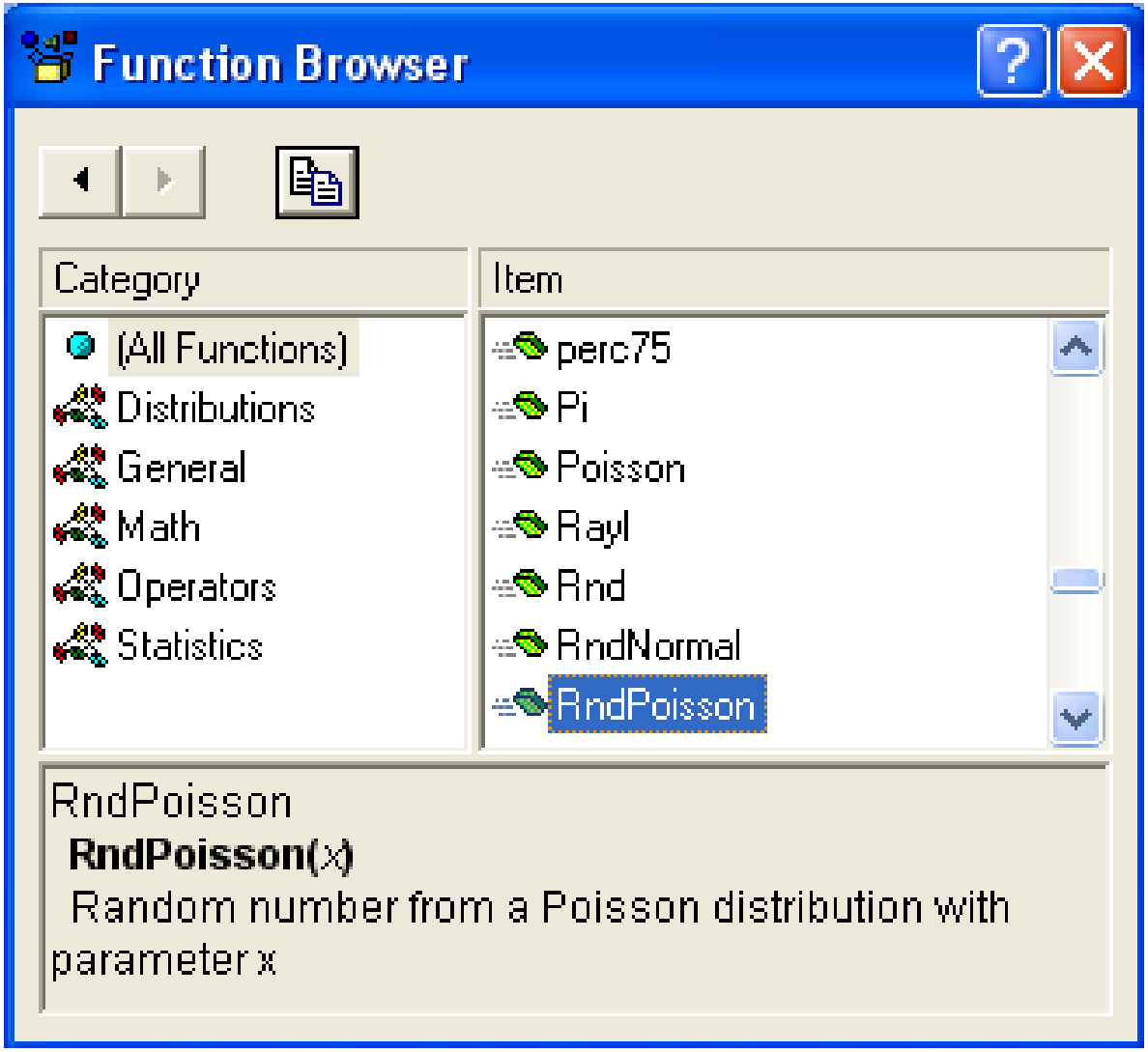

Function Browser
Category
Item
(All Functions)
Distributions
General
Math
Operators
Statistics
perc75
Pi
Poisson
Rayl
Rnd
RndNormal
RndPoisson
RndPoisson
RndPoisson(x)
Random number from a Poisson distribution with parameter x

Рис.15