Gnumeric: электронная таблица для всех

И.А.Хахаев, © 2007-2010

5 Инструменты Gnumeric для статистиков

Инструменты статистической обработки данных находятся в пункте главного меню «Сервис/Статистический анализ» (рис. 5.1). В этой главе рассмотрим принципы работы большинства из них, поскольку от версии к версии добавляются новые инструменты и возможности.

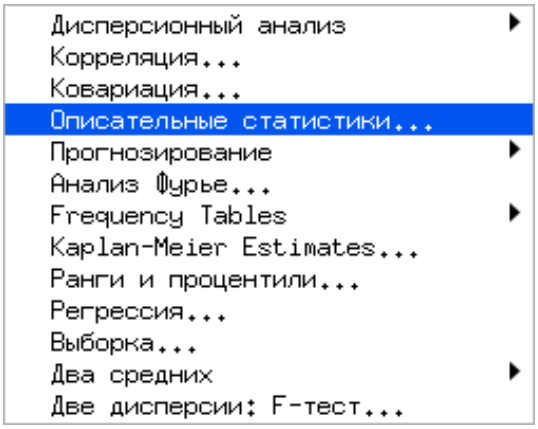

Рисунок 5.1. Средства статистического анализа

5.1 Описательные статистики

Исследование возможностей Gnumeric по статистической обработке данных начнем с простейшей задачи – получения основных статистических характеристик выборки. В качестве исходных данных будем использовать диапазоны ячеек, заполненные последовательностью случайных чисел. Примеры данных здесь приводить не имеет смысла, поэтому будет описываться вид исходного модельного распределения и его параметры, а на рисунках будут приводиться диалоги формирования исходных данных и результаты.

Для начала сформируем выборку с нормальным распределением, задав среднее значение 5 и стандартное отклонение 1. Это делается в помощью диалога «Генерация случайных чисел» (рис. 5.2), вызываемого из главного меню («Правка/Заполнить/Генерация случайных чисел...» или «Данные/Заполнить/Генерация случайных чисел...»).

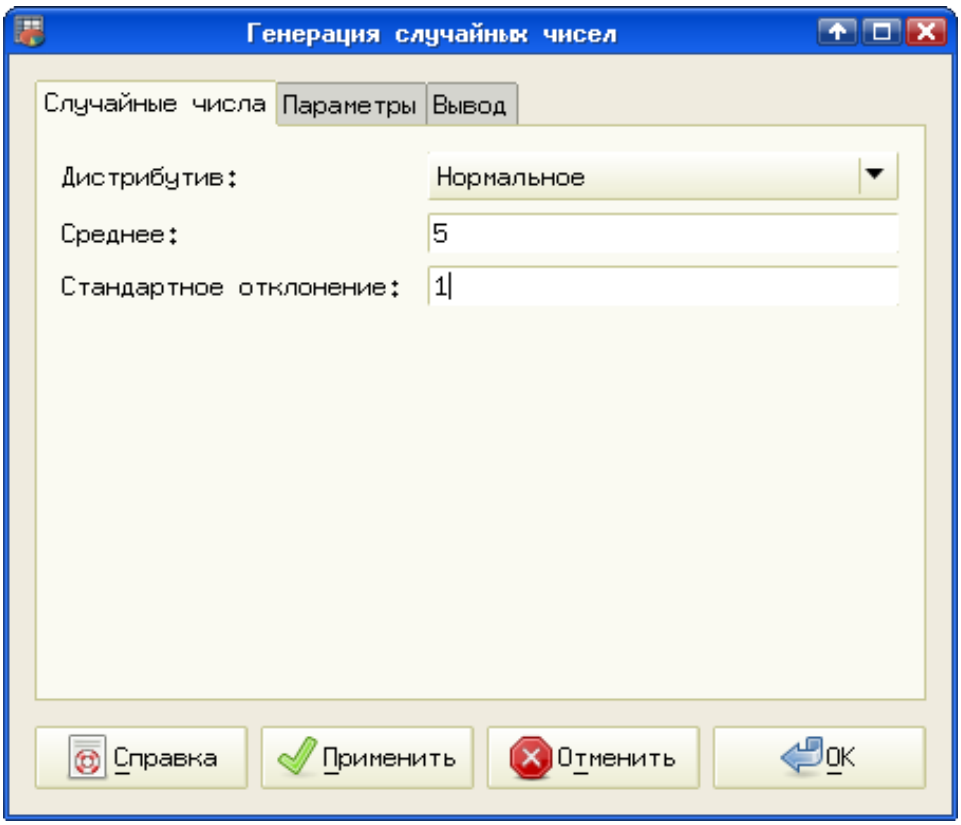

Рисунок 5.2. Выбор распределения для создания исходных данных

На вкладке «Случайные числа» устанавливаем вид распределения (почему-то названный «Дистрибутив») - Нормальное, Среднее значение – 5 и Стандартное отклонение – 1. На вкладке «Параметры» устанавливаем Число переменных – 1 и Размер выборки – 25 (рис. 5.3).

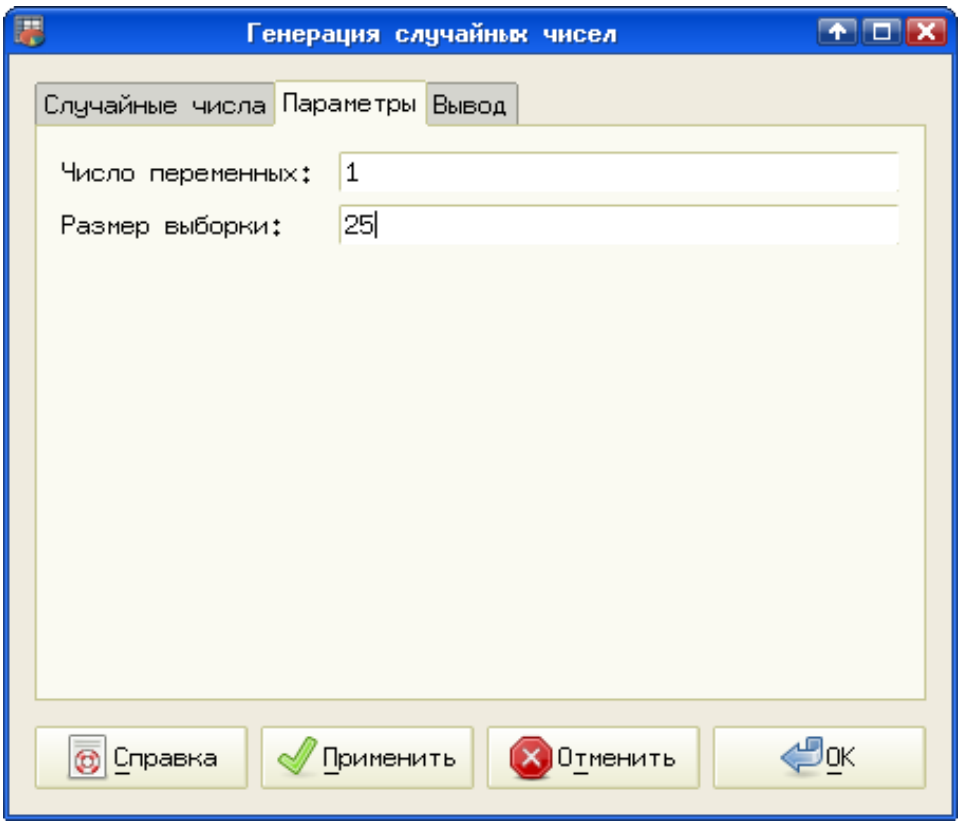

Рисунок 5.3. Определение параметров выборки

Наконец, на вкладке «Вывод» устанавливаем диапазон вывода – диапазон ячеек, начиная, например, с A4 на текущем листе (рис. 5.4). После нажатия на кнопки «Применить» и «ОК» будет получено 25 случайных чисел с заданным законом распределения.

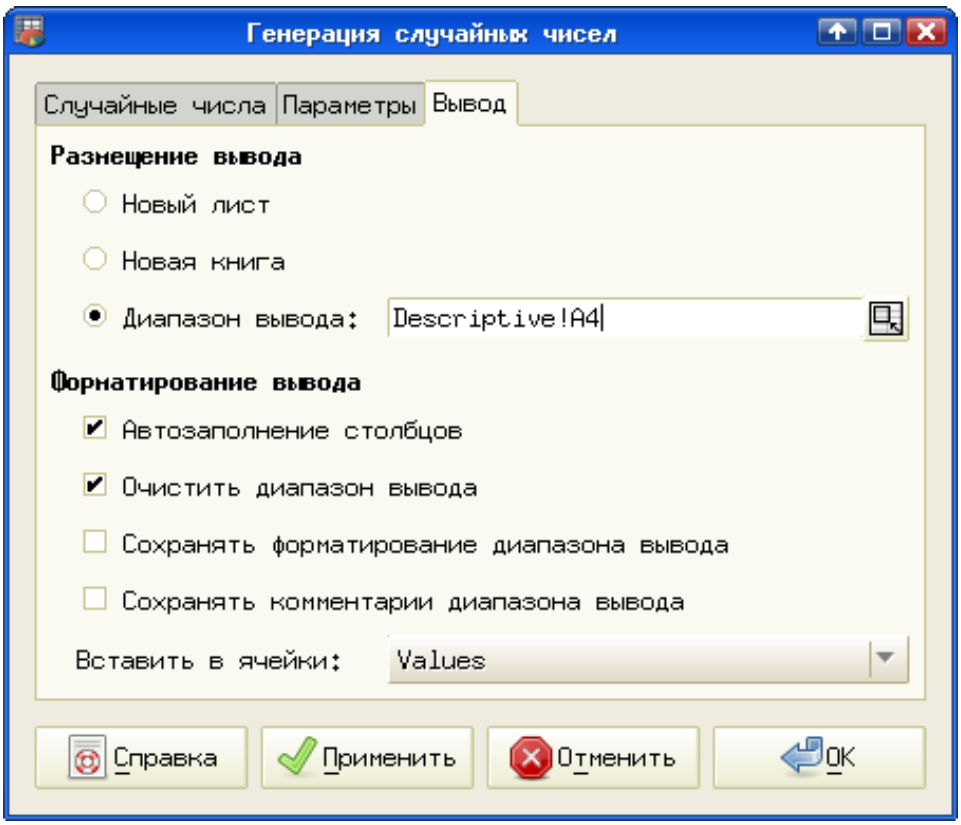

Рисунок 5.4. Настройка размещения результатов

Теперь получим базовые статистические характеристики этой выборки как будто мы про неё ничего не знаем. Для этого выделим наши данные и вызовем диалог «Описательные статистики» («Сервис/Статистический анализ/Описательные статистики...») (рис. 5.5).

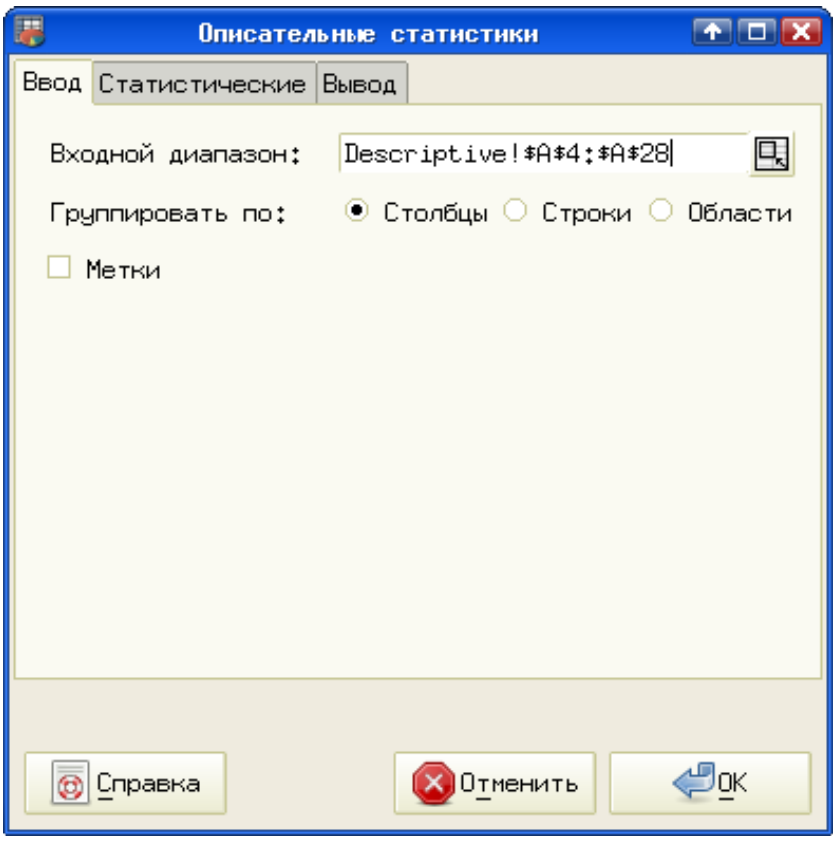

Рисунок 5.5. Определение диапазона данных для обработки

На вкладке «Ввод» проверяем правильность диапазона ввода, на вкладке «Статистические» при необходимости уточняем доверительный интервал и другие параметры (можно все оставить по умолчанию, как на рис. 5.6). Наконец, на вкладке «Вывод» опять-таки задаем ячейку текущего листа, с которой начнется вывод результатов (рис. 5.7).

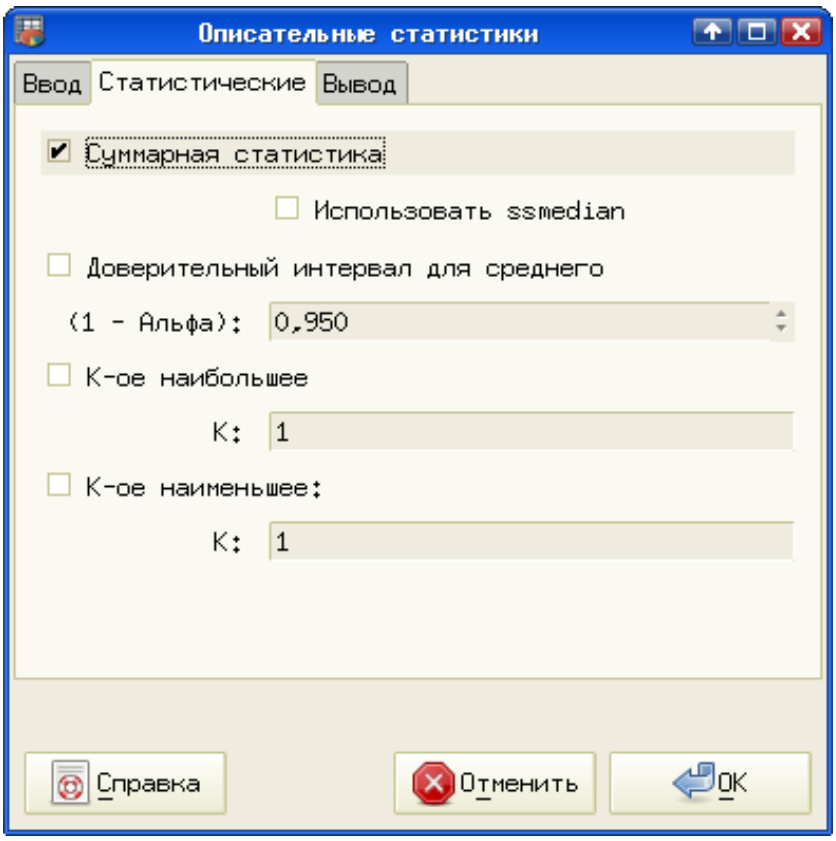

Рисунок 5.6. Уточнение параметров обработки данных

Очевидно, что раз 25 точек данных начинаются с A4, то имеет смысл выводить результаты статистического анализа после окончания данных. Хотя ничто не мешает вывести результаты на отдельный лист, это не очень удобно, если приходится сравнивать характеристики нескольких выборок.

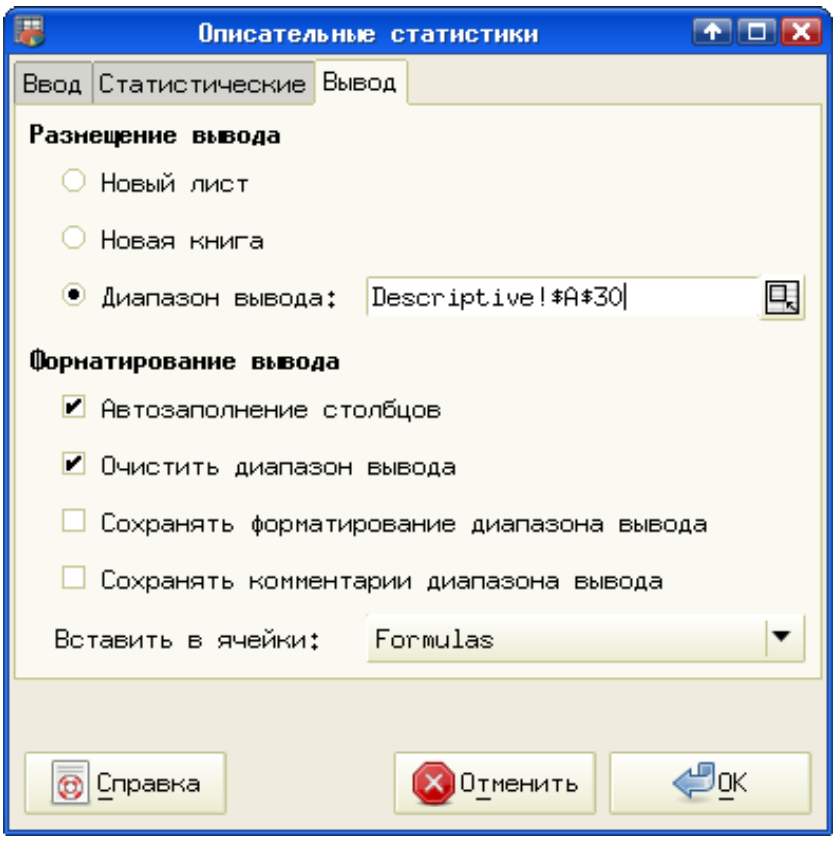

Рисунок 5.7. Определение размещения результатов обработки

В дальнейшем без особой необходимости все эти однотипные диалоги приводиться не будут. Теперь посмотрим на результаты обработки исходных данных – те самые описательные статистики для нормального распределения (рис. 5.8).

30		*Столбец 1*
31	Среднее	5,04464909892853
32	Стандартная ошибка	0,18236526945984
33	Медиана	5,01330411151615
34	Мода	#N/A
35	Стандартное отклонение	0,91182634729922
36	Выборочная дисперсия	0,83142728762905
37	Эксцесс	1,11813424036683
38	Ассиметрия	0,65839749421218
39	Диапазон	4,06723579731569
40	Минимум	3,45842165522619
41	Максимум	7,52565745254188
42	Сумма	126,116227473213
43	Количество	25

Рисунок 5.8. Описательные статистики для нормального распределения

Из приведённых результатов видно, что сгенерированы были действительно случайные числа. Вычисленные по выборке значения близки к параметрам, по

которым формировалась эта выборка, но совпадение не идеальное, т.е. фактор «случайности» действительно имеет место. Отсутствие значения для моды, вероятно, связано с тем, что исходная выборка воспринимается как вариативный ряд, в котором нет варианты с максимальной частотой, поскольку значения не повторяются.

Теперь проделаем те же операции для равномерного распределения в диапазоне [-2;2] и посмотрим на результаты (рис. 5.9).

	Столбец 1
Среднее	0,15410167465159
Стандартная ошибка	0,18644560472611
Медиана	0,15616328055933
Мода	#N/A
Стандартное отклонение	0,93222802363054
Выборочная дисперсия	0,86904908804211
Эксцесс	-0,66484799688632
Ассиметрия	0,10377506362021
Диапазон	3,56023754139636
Минимум	-1,63299821817276
Максимум	1,9272393232236
Сумма	3,8525418662898
Количество	25

Рисунок 5.9. Описательные статистики для равномерного распределения

Здесь стандартное отклонение очень велико по сравнению с диапазоном от минимума до максимума, что неудивительно для равномерного распределения.

Таким образом, инструмент «Описательные статистики» позволяет получить практически все необходимые статистические характеристики имеющейся выборки.

5.2 Прогнозирование

Статистическое прогнозирование (в англоязычных статистических программах – forecasting) является на самом деле сглаживанием, которое применяется для выделения тенденции при сильном разбросе точек исходных данных. В Gnumeric эта процедура может проводиться двумя способами – методом экспоненциального сглаживания и методом скользящего среднего (соответственно, команды главного меню «Сервис/Статистический анализ/Прогнозирование/Экспоненциальное сглаживание...» и «Сервис/Статистический анализ/Прогнозирование/Скользящее среднее...»). При выборе сглаживания методом скользящего среднего можно указать количество точек, по которым будет проводиться усреднение.

Рассмотрим пример с некоторыми экспериментальными данными (рис. 5.10). Вектор X представляет собой некоторую независимую переменную, вектор Y – измеренные значения. Правее приведены результаты экспоненциального сглаживания и сглаживания методом скользящего среднего по трем точкам. Поскольку при сглаживании для данного значения Y оказываются задействованы предыдущие и последующие значения, то количество «сглаженных» точек меньше,

чем количество исходных. Это видно как по отсутствию последнего значения в обоих случаях сглаживания, так и из сообщения «#N/A (нет данных)» в начале последовательности. Для скользящего среднего по трем точкам результат вообще начинается только с третьей точки последовательности.

2				
3			Эксп. сглаживание	Скольз. среднее(3)
4			*Столбец 1*	*Столбец 1*
5	X	Y	#N/A	#N/A
6	1	80,54	80,540000	#N/A
7	2	54,21	59,476000	61,920000
8	3	51,01	52,703200	43,493333
9	4	25,26	30,748640	31,566667
10	5	18,43	20,893728	18,933333
11	6	13,11	14,666746	14,763333
12	7	12,75	13,133349	11,643333
13	8	9,07	9,882670	9,406667
14	9	6,4	7,096534	6,633333
15	10	4,43	4,963307	4,740000
16	11	3,39	3,704661	3,326667
17	12	2,16	2,468932	2,416667
18	13	1,7	1,853786	1,666667
19	14	1,14	1,282757	1,163333
20	15	0,65		

Рисунок 5.10. Исходные данные и результаты сглаживания

График исходных данных и результатов сглаживания показан на рис. 5.11. Нужно заметить, что сами операции сглаживания («прогнозирования») дают только числовые значения.

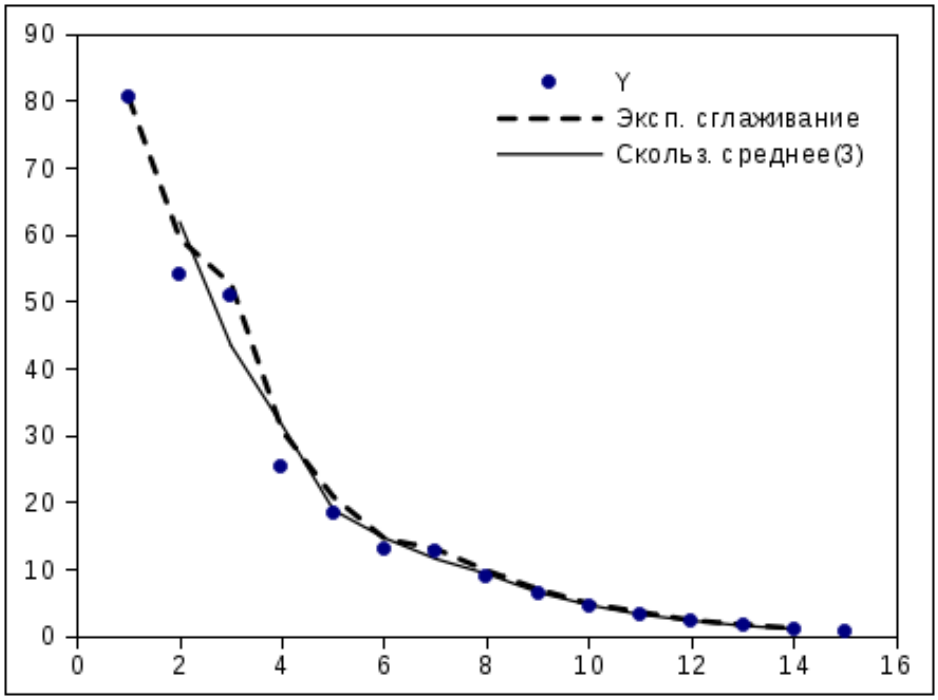

Рисунок 5.11. Графическое представление исходных данных и результатов сглаживания

Из графика видно, что скользящее среднее (сплошная линия) в данном примере дает лучший результат, но чем больше точек участвуют в усреднении (и чем более гладкой получается кривая), тем больше точек в начале и в конце теряются.

5.3 Корреляция

Корреляционный анализ, как известно, позволяет выявить взаимосвязь двух случайных величин. Коэффициент корреляции (корреляция по Пирсону) может принимать значения от -1 до 1. Чем ближе к 1 абсолютная величина коэффициента корреляции, тем сильнее связаны исследуемые случайные величины.

Для примера рассмотрим несколько выборок. Первая выборка (назовем ее «Выборка1») является нормально распределенной случайной величиной со средним значением 5 и стандартным отклонением 1, вторая («Выборка2») - также нормально распределенная случайная величина со средним значением 4 и стандартным отклонением 2, третья («Выборка3») - случайная величина с равномерным распределением в интервале [-2;2]. Четвертая выборка («Выборка4») получена путем удвоения значений Выборки1 и добавления к результату значений Выборки3 (т.е. Выборка4=2*Выборка1+Выборка3) в каждой точке.

Пусть в каждой выборке будет по 25 значений, и начинать генерацию исходных данных будем со столбца A. После создания выборок вызываем диалог «Корреляция» («Сервис/Статистический анализ/Корреляция...», рис. 5.12).

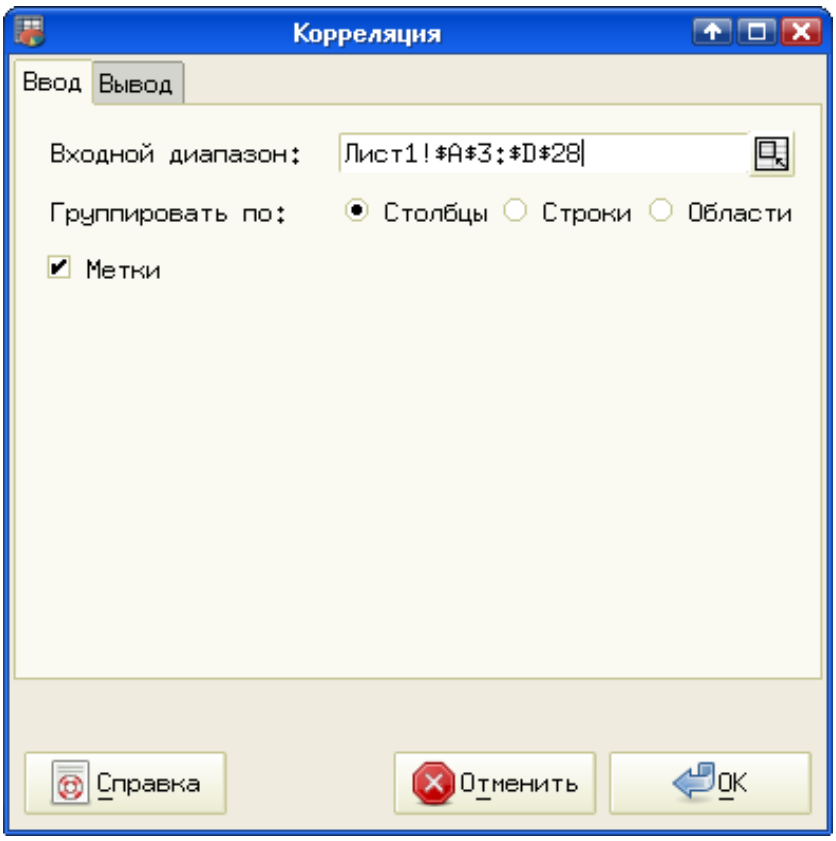

Рисунок 5.12. Определение исходных данных для вычисления корреляции

Присвоение имён столбцам исходных данных и использование режима «Метки» в диалоге «Корреляция» позволяет на выходе получить именованные результаты (рис. 5.13).

30					
31	*Корреляции*	*Выборка1*	*Выборка2*	*Выборка3*	*Выборка4*
32	*Выборка1*	1,000000			
33	*Выборка2*	-0,330514	1,000000		
34	*Выборка3*	-0,282875	0,276372	1,000000	
35	*Выборка4*	0,999637	-0,325260	-0,256946	1,000000

Рисунок 5.13. Результаты вычисления корреляции выборок

Видно, что каждая выборка сама с собой прекрасно коррелирует (коэффициент равен 1), а четвертая выборка с первой дает коэффициент почти равный 1 (почти, но не совсем, поскольку Выборка4 искажена дополнительным влиянием Выборки3).

Для упражнения полезно вычислить коэффициент корреляции двух независимых выборок случайных величин с одинаковыми параметрами распределения.

5.4 Ковариация

Коэффициент ковариации также позволяет определить взаимосвязи случайных величин, но, в отличие от коэффициента корреляции, этот параметр не является нормированным, поэтому его значение не несет никакой очевидной информации. Более информативным является вычисление коэффициента корреляции. Посмотрим на результаты вычисления ковариации («Сервис/Статистический анализ/Ковариация...») для тех же исходных данных, что и в примере вычисления корреляции (рис. 5.14).

37	Ковариации	Выборка1	Выборка2	Выборка3	Выборка4
38	Выборка1	3,573914			
39	Выборка2	-1,428732	5,22851		
40	Выборка3	-0,563336	0,66571	1,109692	
41	Выборка4	7,091494	-2,79089	-1,015704	14,081418

Рисунок 5.14. Результат вычисления ковариации выборок

Видно, что без дополнительных усилий как-то интерпретировать результаты вычисления ковариации достаточно сложно.

5.5 Регрессия

Регрессия как элемент статистического анализа в Gnumeric проводится по линейной модели, а отклонения рассматриваются как нормально распределённые случайные ошибки. В результате проведения такого регрессионного анализа («Сервис/Статистический анализ/Регрессия...») вычисляется множество параметров, которые могут многое сказать опытному специалисту.

Рассмотрим результаты регрессии для примера исходных данных, использованных ранее для сглаживания (см. рис. 5.10).

Итоговый вывод			
Регрессионные статистики			
Множественная регрессия	-0,852668		
Коэффициент определенности	0,727043		
Подобранный коэффициент определенности	0,706046		
Стандартная ошибка	13,054635		
Наблюдения	15		
Дисперсионный анализ			
	степень свободы	сумма квадратов	Квадрат среднего
Регрессия	1	5901,17958892857	5901,17958892857
Остатки	13	2215,50541107143	170,423493159341
Всего	14	8116,685	
	F	Значимость F	
	34,626561629101	5,37181850557115E-05	
	Коэффициенты	Стандартная ошибка	t Stat
Пересечение	55,676571428571	7,09334247757616	7,8491305903499
Столбец 1	-4,590821428571	0,78016365764622	-5,88443384099957
	Значение P	Ниже 95%	Выше 95%
	2,752699921E-06	40,3523366704577	71,0008061866851
	5,371818506E-05	-6,27626254146192	-2,90538031568094

Рисунок 5.15. Результаты регрессионного анализа экспериментальных данных

При создании иллюстрации (рис. 5.15) данные из длинных строчек (параметры «F» и «P») были перенесены вниз, т.е. в реальной таблице ячейки «F» и «Значимость F» находятся в той же строке, что и параметр «степень свободы», а значения параметров «P», «Ниже 95%» и «Выше 95%» - в тех же строках, что и значения коэффициентов регрессии.

В результате получается уравнение Y=-4.59*X+55.677. Параметр «Столбец1» дает коэффициент наклона прямой, а параметр «Пересечение» - точку пересечения прямой с осью Y.

Интересно сравнить результаты регрессионного анализа, проведенного таким образом, с уравнениями регрессии, которые можно получить на диаграмме XY (рис. 5.16).

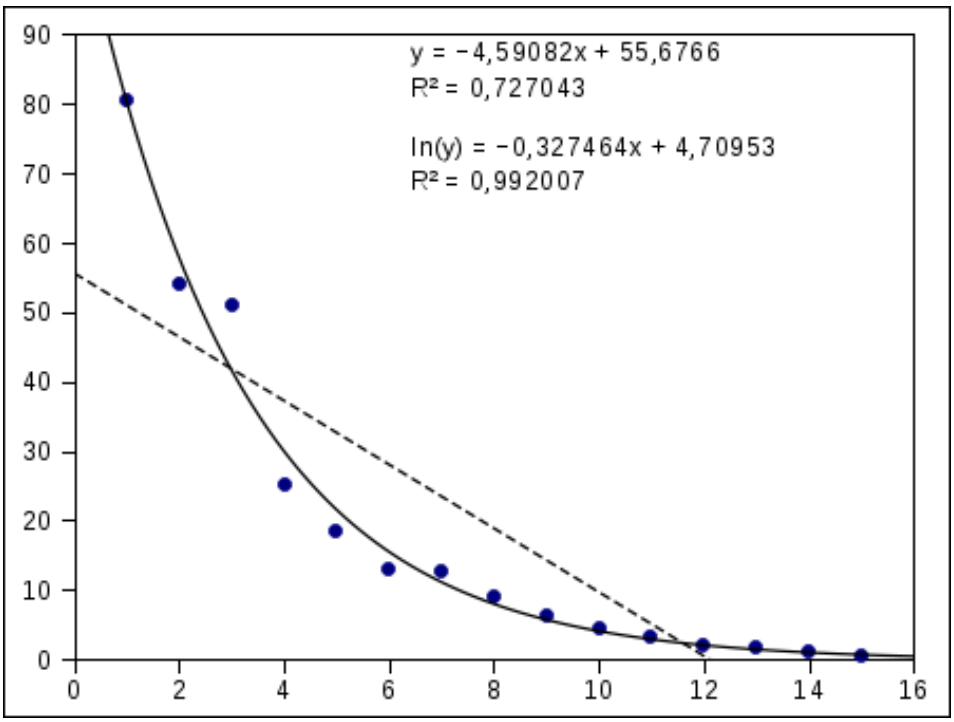

Рисунок 5.16. Исходные данные и кривые регрессии на диаграмме XY

На рис. 5.16 точками показаны исходные данные, пунктирной линией – линейная регрессия (верхнее уравнение), параметры которой в точности совпадают с вычисленными с помощью «статистического» регрессионного анализа. Сплошная линия и нижнее уравнение соответствуют экспоненциальной модели регрессии, которая дает гораздо лучший коэффициент определенности (критерий Пирсона).

Более подробно построение кривых регрессии будет рассматриваться в главе «Регрессионный анализ в Gnumeric».

5.6 Анализ Фурье

Модуль преобразования Фурье («Сервис/Статистический анализ/Анализ Фурье...») позволяет вычислять дискретное преобразование Фурье (ДПФ) для заданного ряда данных (режим «Инверсия» на вкладке «Параметры» диалога «Анализ Фурье» позволяет, по-видимому, вычислять обратное преобразование Фурье).

Рассмотрим два примера. Первый пример – преобразование Фурье для гауссовой кривой, задаваемой формулой (1).

$$y = e^{\left(-\frac{(x-x_0)^2}{\sigma^2}\right)} \qquad (1)$$

Известно, что для такой функции Фурье-образ будет иметь вид такой же функции (с точностью до нормировки). Путь для простоты x_0 имеет значение 0, а дисперсия пусть будет равна 4. Графики исходной функции (пунктир) и ее Фурье-

образа (сплошная линия) показаны на рис. 5.17.

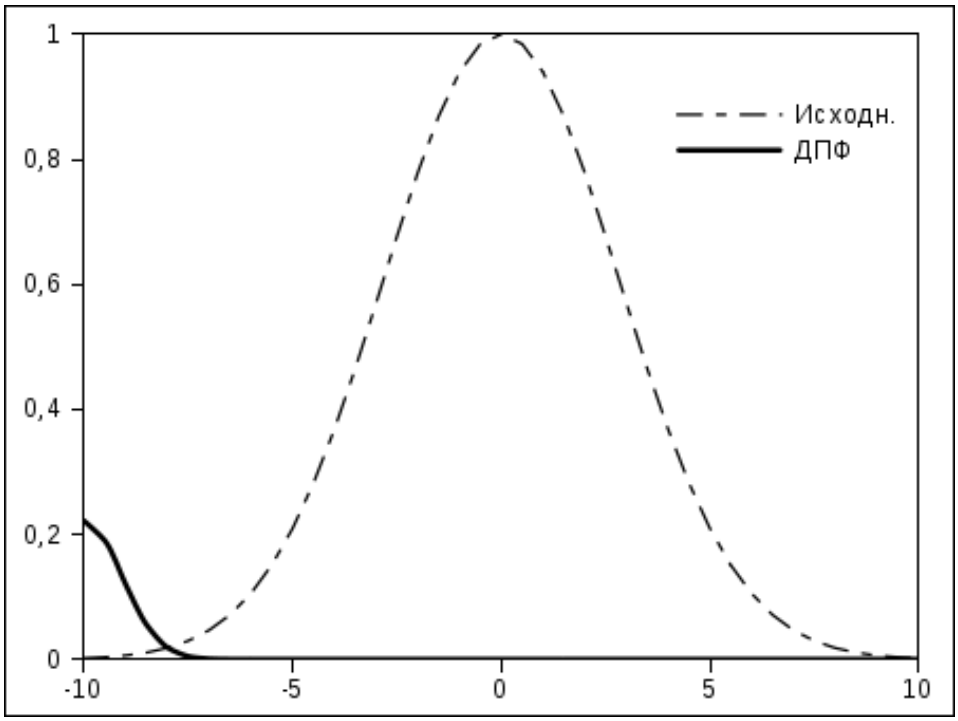

Рисунок 5.17. Гауссова кривая и её Фурье-образ

В Gnumeric вычисляются действительная и мнимая части Фурье-образа, так что для получения окончательного результата можно воспользоваться математической функцией `hypot()`, вычисляющей квадратный корень из суммы квадратов аргументов.

Теперь попробуем получить Фурье-образ для периодической «прямоугольной» функции (прямоугольных импульсов), причем состояние «0» и состояние «1» длятся по половине периода (рис. 5.18).

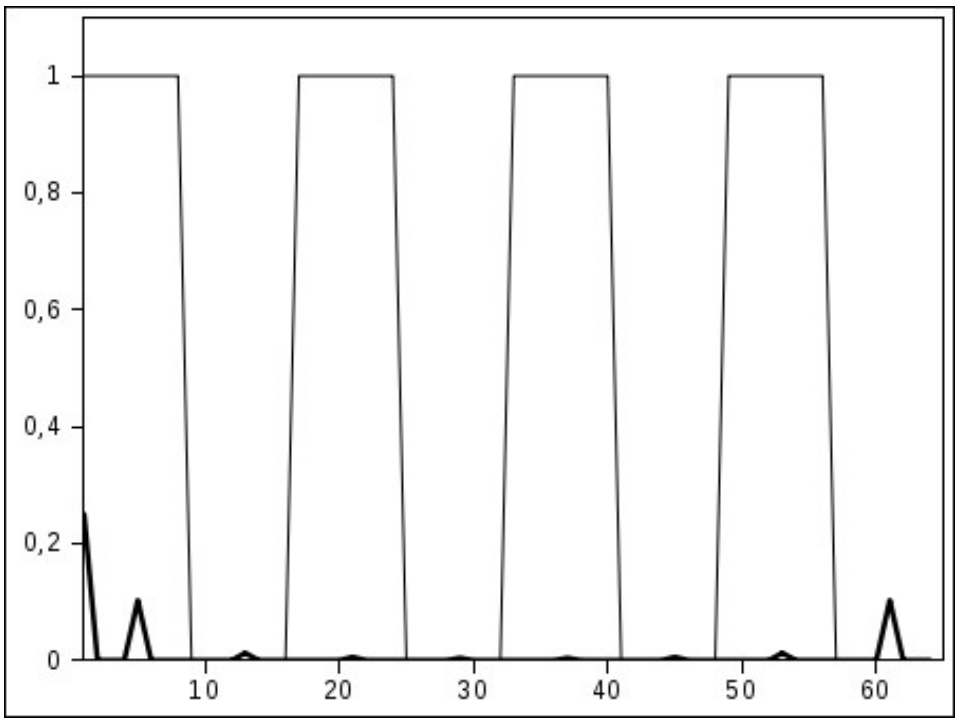

Рисунок 5.18. «Импульсная» функция и её Фурье-образ

По горизонтальной оси отложены некие условные единицы. Период исходной функции составляет 16 единиц (всего взято 64 точки). Фурье-образ показан более жирной линией. Видно, что Фурье-образ симметричен относительно середины горизонтальной оси. Чем больше точек по горизонтали взято и чем больше период исходной функции, тем больше компонентов ряда Фурье можно различить.

Специально для читателей с инженерным и техническим образованием нужно заметить, что на рис. 5.18 показан квадрат модуля Фурье-образа, что соответствует, например, распределению энергии при дифракции электромагнитных волн на периодической решетке. В полном соответствии с теорией дифракции для заданных параметров решетки «энергия» в первом порядке (первый слева пик) составляет около 10% абсолютной величины.

5.7 Гистограмма

Инструмент «Гистограмма» («Сервис/Статистический анализ/Частотная таблица/Гистограмма...» или «Сервис/Статистический анализ/Frequency Tables/Гистограмма...» в зависимости от версии) вычисляет количество значений в выборке, попадающих в заданный интервал значений. Границы интервалов (отрезки, cutoffs) могут быть заданы заранее или вычислены исходя их максимального и минимального значений и желаемого количества интервалов (рис. 5.19).

В качестве тестовых значений сформируем выборку из 39 нормально распределённых случайных величин со средним значением 5 и стандартным отклонением 2.

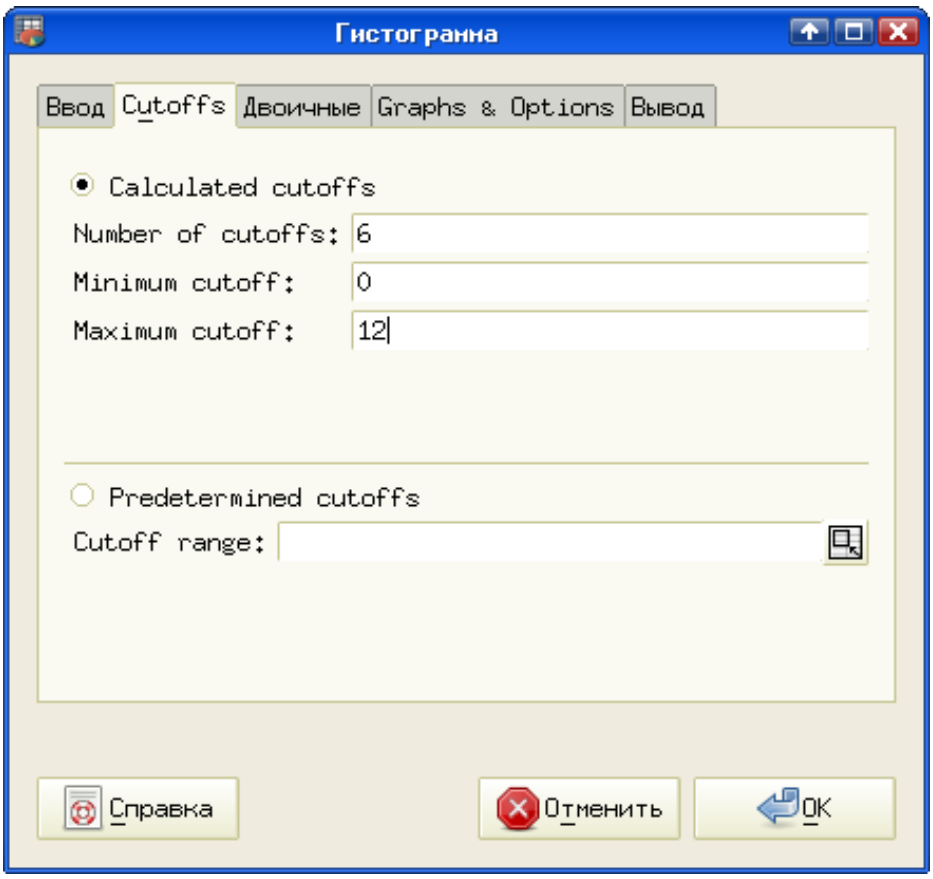

Рисунок 5.19. Определение диапазонов значений для создания гистограммы

(На вкладке «Ввод» обычным образом задаётся диапазон ячеек с исходными данными, поэтому эту вкладку диалога настройки гистограммы не обсуждаем).

На вкладке «Двоичные» определяется способ учёта значений, на границах отрезков (рис. 5.20). Если какое-то значение точно (с учётом «машинного нуля») попадает на границу интервала (отрезка), то для границы, отмеченной квадратной скобкой («[» или «]») оно учитывается в этом интервале (отрезке), а для границы, отмеченной круглой скобкой – в соседнем (предыдущем или следующем).

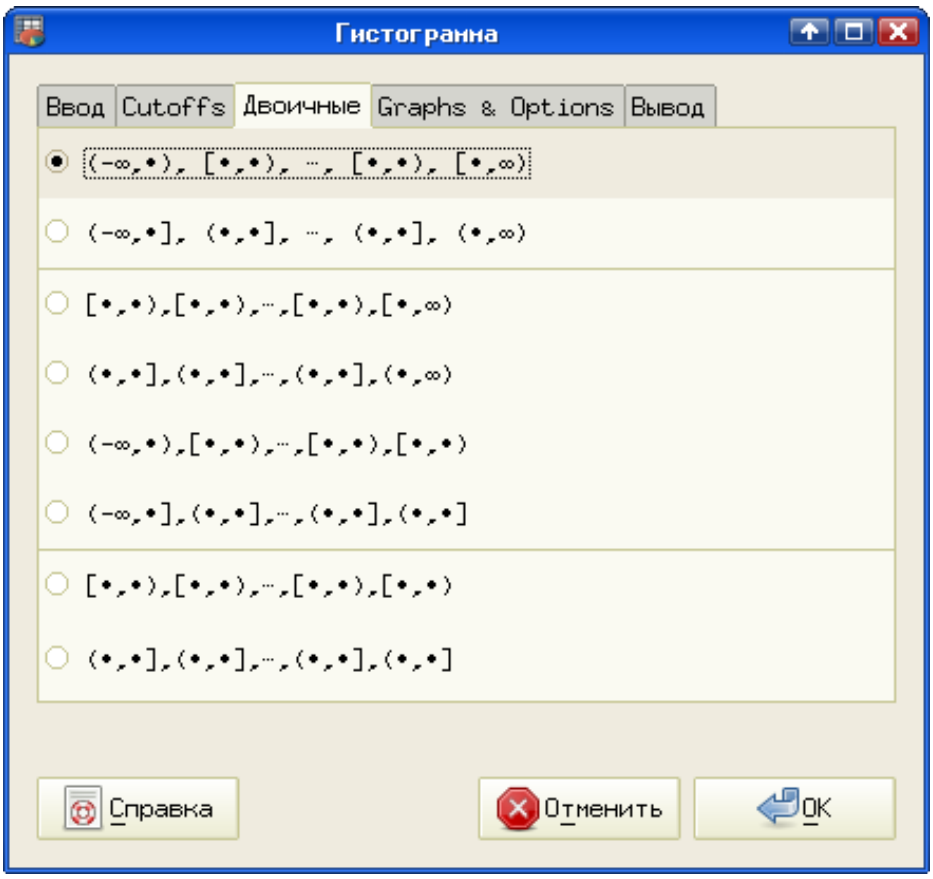

Рисунок 5.20. Определение правил учёта границ отрезков

На вкладке «Графики и параметры» (Graphs & Options) нужно определить вид диаграммы, которая будет сформирована и формат вывода результатов. Достаточно разумно заказать вывод гистограммы и представления результатов в процентах, как показано на рис. 5.21.

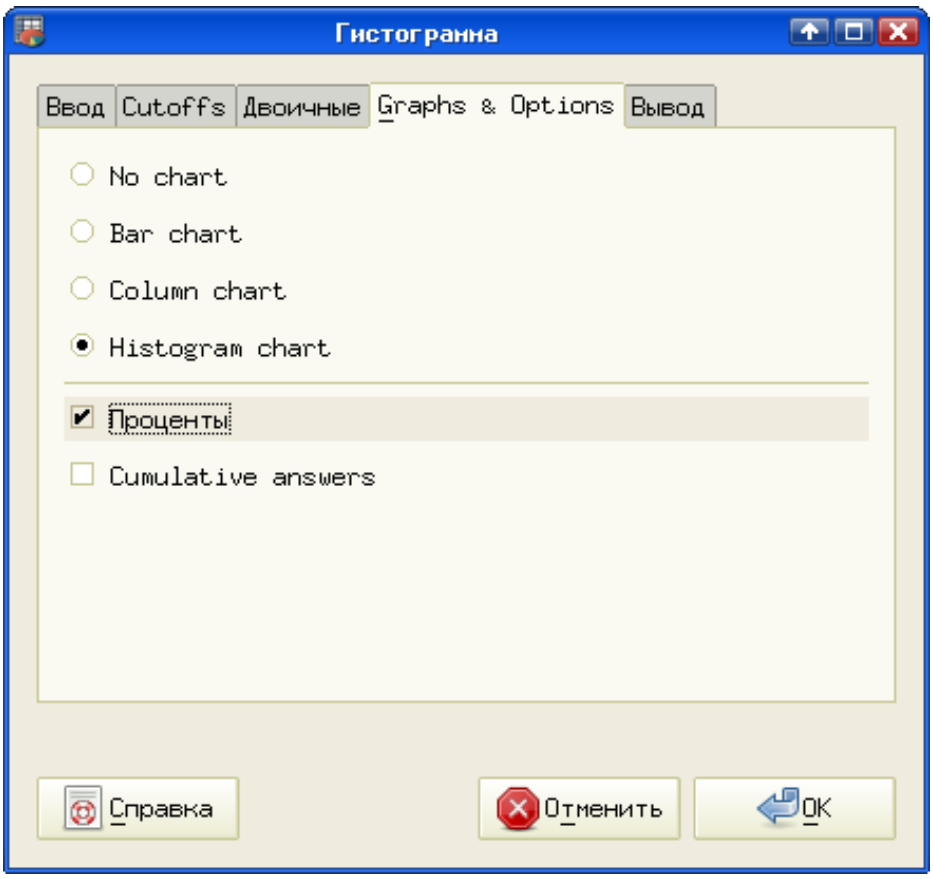

Рисунок 5.21. Определение вида получаемой диаграммы и формата вывода результатов

Наконец, на вкладке «Вывод», как обычно в Gnumeric, определяется лист и диапазон ячеек на листе, в который будут выводиться результаты (рис. 5.22).

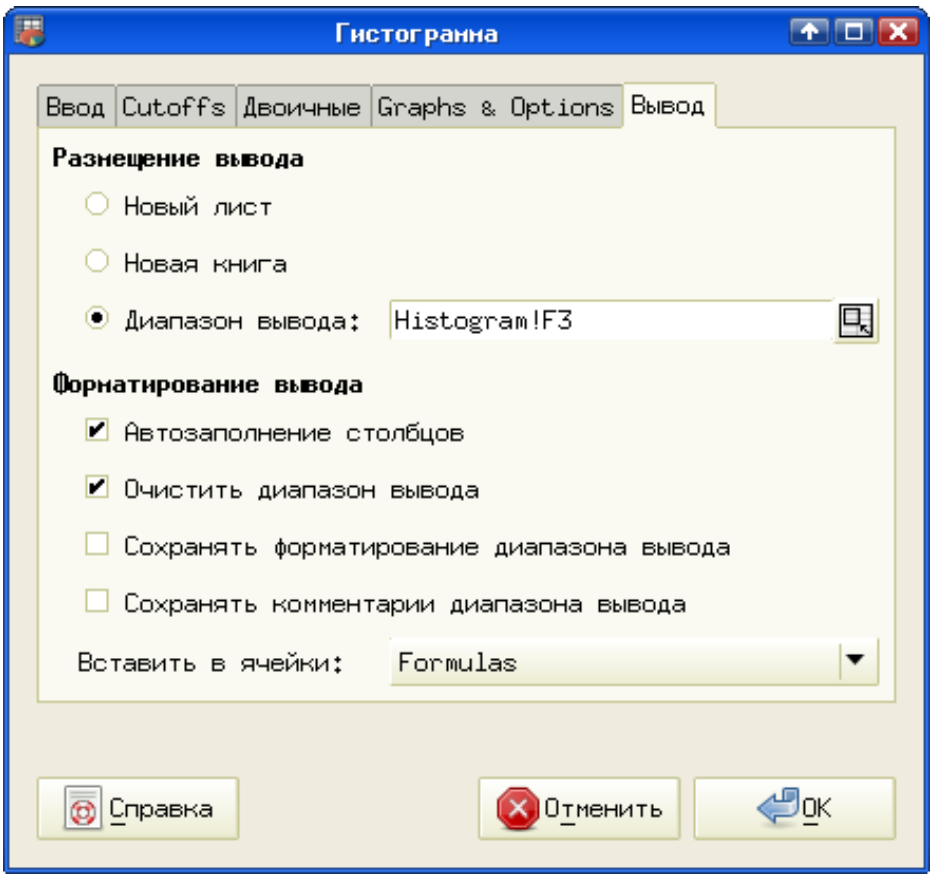

Рисунок 5.22. Указание диапазона для вывода результатов

После нажатия на кнопку «ОК» строится гистограмма и вычисляются частоты попадания значений выборки в заданные отрезки. Однако позиция графика гистограммы и диапазона ячеек с результатами совпадают, поэтому график нужно отодвинуть, чтобы увидеть числа (см. рис. 5.23).

Гистограмма

			Столбец 1
from	-∞	to below 0	0,0%
from	0	to below 2,4	15,0%
from	2,4	to below 4,8	47,5%
from	4,8	to below 7,2	32,5%
from	7,2	to below 9,6	5,0%
from	9,6	to below 12	0,0%
from	12	to	∞

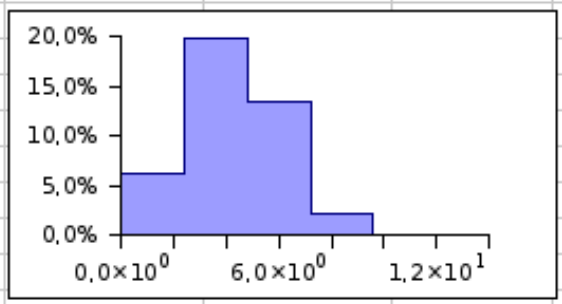

Рисунок 5.23. Результаты вычислений и график гистограммы

График гистограммы можно теперь настраивать обычным образом, изменяя размер, форматы вывода по осям, цвет для серии данных и т. д.

5.8 Выборка

С помощью диалога «Выборка» можно выбрать несколько серий данных из некоторого вектора данных. Для примера в качестве исходного вектора рассмотрим столбец чисел от 0 до 25 (назовем столбец словом «Данные»). В диалоге «Выборка» («Сервис/Статистический анализ/Выборка...») на вкладке «Ввод», как обычно, определяется диапазон ячеек, содержащих исходные данные, на вкладке «Вывод» - расположение результатов.

Вкладка «Параметры» (рис. 5.24) заслуживает отдельного обсуждения.

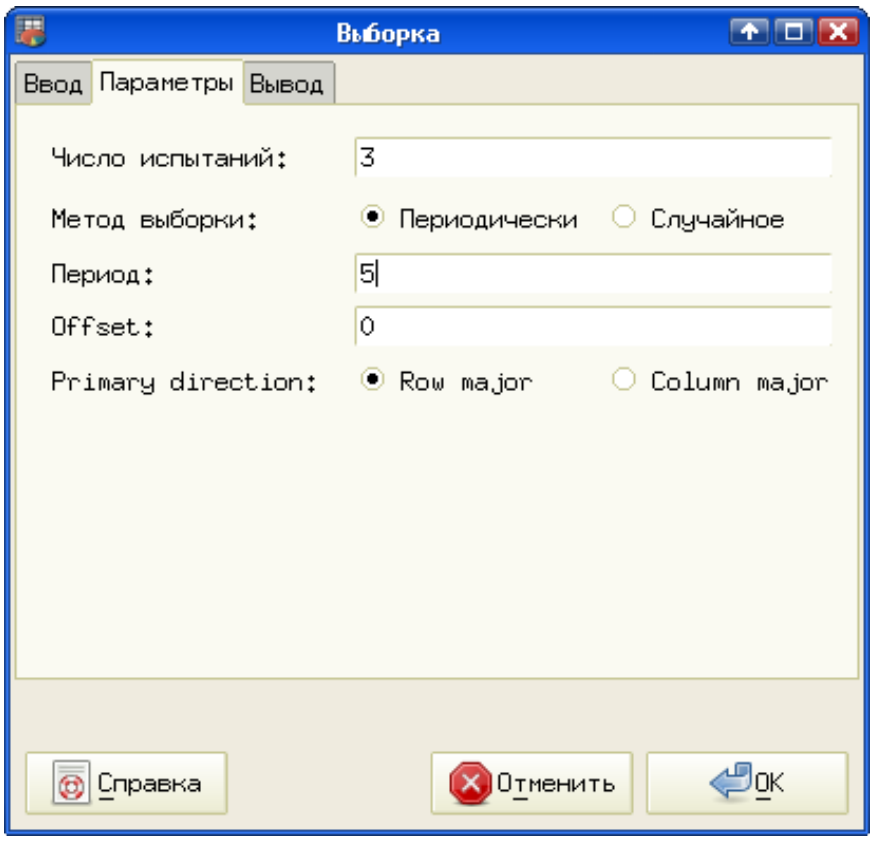

Рисунок 5.24. Определение параметров периодической выборки

Есть возможность задать периодическую выборку и случайную. Пример параметров для периодической выборки показан на рис. 5.24. Параметр «Offset» («Смещение») определяет начальную позицию в векторе исходных данных. Для вектора, имеющего имя, его можно оставить равным 0, если на вкладке «Ввод» установлен режим «Метки» (всё становится понятным, если провести несколько экспериментов). Результат такой выборки показан на рис. 5.25.

Данные	*Данные*	*Данные*
4	4	4
9	9	9
14	14	14
19	19	19
24	24	24

Рисунок 5.25. Результат трёх испытаний для периодической выборки

Видно, что при периодической выборке размер периода задаёт интервал между выбираемыми значениями (в данном примере выбирается каждое 5-е значение, начиная с номера 0), и в различных сериях (испытаниях) значения повторяются.

Пример параметров случайной выборки показан на рис. 5.26, а результат – на рис. 5.27.

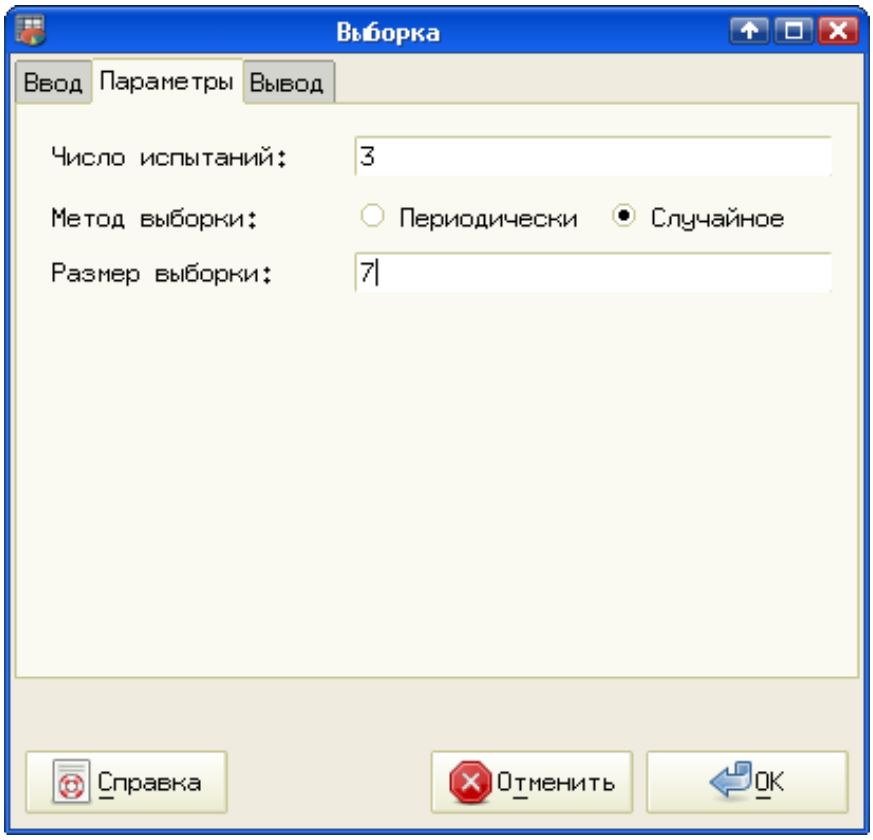

Рисунок 5.26. Определение параметров случайной выборки

Данные	Данные	Данные
8	24	6
24	7	4
19	9	1
17	17	13
1	21	17
3	4	20
0	10	14

Рисунок 5.27. Результат трёх испытаний при случайной выборке

Видно, что для случайной выборки количество точек определяется параметром «Размер выборки» и в различных сериях (испытаниях) значения не повторяются.

Инструмент «Выборка» может быть использован для превращения вектора данных в матрицу, а также для «прореживания» больших наборов исходных данных.

5.9 Ранги и процентили

Инструмент «Ранги и процентили» выполняет сортировку выборки по убыванию, определяет максимальное значение в выборке, присваивая ему первый ранг и значение в 100%, остальные значения выстраивает относительно первого и определяет, какой процент составляет текущая величина в выборке относительно максимального значения, а также указывает позицию каждого значения в выборке.

Рассмотрим, например, гипотетические результаты единого госэкзамена по ботанике среди группы в 25 участников (фамилии не указаны, поскольку они не используются для формирования данных). Таблица исходных данных приведена на рис. 5.28 (столбец «Баллы»).

	Средний ранг				Верхний ранг			
Баллы	Точки	Баллы	Ранг	Процентиль	Точки	Баллы	Ранг	Процентиль
12	7	93	1,5	100,00%	7	93	1	100,00%
34	10	93	1,5	100,00%	10	93	1	100,00%
28	6	88	3	91,67%	6	88	3	91,67%
47	22	78	4	87,50%	22	78	4	87,50%
67	14	77	5,5	83,33%	14	77	5	83,33%
88	23	77	5,5	83,33%	23	77	5	83,33%
93	15	68	7	75,00%	15	68	7	75,00%
45	5	67	8,5	70,83%	5	67	8	70,83%
48	20	67	8,5	70,83%	20	67	8	70,83%
93	11	65	10	62,50%	11	65	10	62,50%
65	16	58	11	58,33%	16	58	11	58,33%
38	9	48	12	54,17%	9	48	12	54,17%
19	4	47	13,5	50,00%	4	47	13	50,00%
77	25	47	13,5	50,00%	25	47	13	50,00%
68	17	46	15,5	41,67%	17	46	15	41,67%
58	24	46	15,5	41,67%	24	46	15	41,67%
46	8	45	17	33,33%	8	45	17	33,33%
43	18	43	18	29,17%	18	43	18	29,17%
34	12	38	19	25,00%	12	38	19	25,00%
67	2	34	21	20,83%	2	34	20	20,83%
34	19	34	21	20,83%	19	34	20	20,83%
78	21	34	21	20,83%	21	34	20	20,83%
77	3	28	23	8,33%	3	28	23	8,33%
46	13	19	24	4,17%	13	19	24	4,17%
47	1	12	25	0,00%	1	12	25	0,00%

Рисунок 5.28. Исходные данные и расчёты рангов и процентилей

На вкладке «Параметры» диалога «Ранги и процентили» (рис. 5.29) можно определить способ вычисления ранга (варианты «Средний ранг» и «Верхний ранг»).

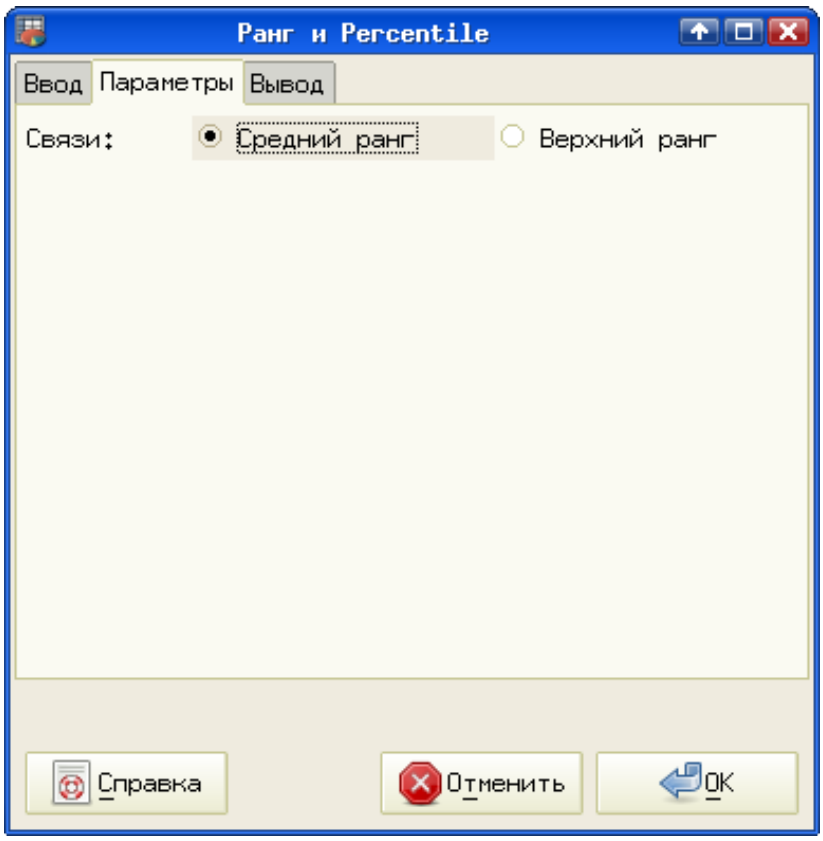

Рисунок 5.29. Определения способа вычисления рангов

Поскольку требуется распределить весь набор значений на количество мест, равных количеству значений, то для повторяющихся значений получается деление мест (два одинаковых значения занимают два уровня рангов). Поэтому при использовании режима «Средний ранг» получаем дробные значения рангов (например, от 13 до 14 ранга имеется два значения, итого в среднем ранг получается 13,5). Результаты вычислений в таком режиме иллюстрируются группой «Средний ранг» на рис. 5.28.

При использовании режима «Верхний ранг» используется минимальное значение ранга для повторяющихся значений, что иллюстрируется группой «Верхний ранг» на рис. 5.28.

5.10 Дисперсионный анализ

Согласно [Гмурман В.Е., Теория вероятностей и математическая статистика, 1977], «дисперсионный анализ (ДА) применяют, чтобы установить, оказывает ли существенное влияние некоторый качественный фактор F, который имеет p уровней F_1, F_2 ... F_p на изучаемую величину X. Основная идея дисперсионного анализа состоит в сравнении «факторной дисперсии», порождаемой воздействие фактора, и «остаточной дисперсии», обусловленной случайными причинами. Если различие этих дисперсий значимо, то фактор оказывает существенное влияние на X ...

... В более сложных случаях исследуют воздействие нескольких факторов на

нескольких постоянных или случайных уровнях и выясняют влияние отдельных уровней и их комбинаций (многофакторный анализ)».

Эта длинная цитата – самое толковое объяснение того, что такое дисперсионный анализ, из всех, которые удалось найти.

- *Однофакторный ДА*

Рассмотрим пример однофакторного ДА на основе данных, взятых из [Бородюк В.П., Вощинин А.П., Иванов А.З и др., Статистические методы в инженерных исследованиях (лабораторный практикум), 1983].

Исследуется зависимость долговечности *y* электрических лампочек (в часах) от технологии изготовления (фактор *x*). В качестве исходных данных используется отклонение долговечности от «стандартного» значения в 1500 часов для 4-х неравночисленных серий образцов из разных партий (см. рис. 5.30).

	y1	y2	y3	y4	y5	y6	y7	y8
серия 1	100	110	150	180	200	200	300	
серия 2	80	140	140	200	250			
серия 3	-40	50	100	120	140	160	240	320
серия 4	10	20	30	70	100	180		

Рисунок 5.30. Исходные данные для однофакторного ДА

В таблице приведены отклонения для различных образцов (y – номера образцов).

Диалог определения исходных данных показан на рис. 5.31.

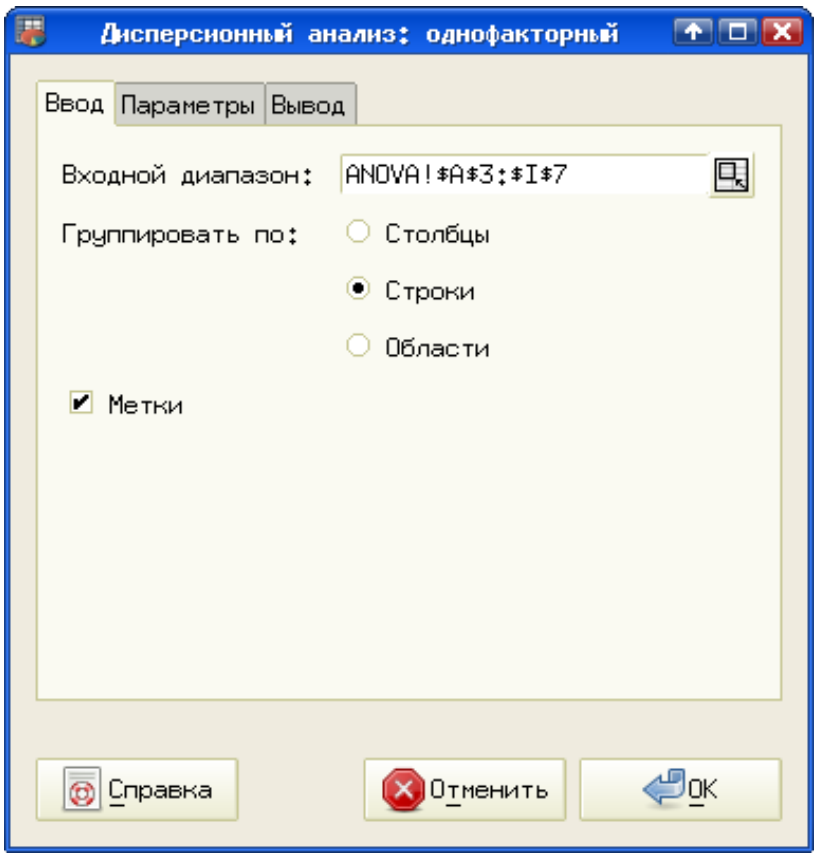

Рисунок 5.31. Определение исходных данных для однофакторного ДА

Важно не включать в диапазон данных лишние ячейки, в данном случае – ячейки с названиями образцов, что обеспечивается включением режима «Метки».

На вкладке «Параметры» устанавливается уровень значимости (по умолчанию – общепринятое значение 5%).

Вкладка «Вывод» - стандартная для всех диалогов статистического анализа, на ней определяется местоположение результатов вычислений.

Результаты показаны на рис. 5.32. Для повышения компактности вывода некоторые ячейки перенесены.

Дисперсионный анализ: однофакторный				
ИТОГО				
Группы	Количество	Сумма	Среднее	Дисперсия
серия 1	7	1240	177,142857142857	4557,1428571429
серия 2	5	810	162	4220
серия 3	8	1090	136,25	12169,642857143
серия 4	6	410	68,3333333333333	4136,6666666667
Дисперсионный анализ				
Источник дисперсии	Сумма квадратов	степень свободы	Квадрат среднего	F
Между группами	42694,771062271	3	14231,5903540904	2,0859969982526
В группах	150093,69047619	22	6822,44047619048	
Всего	192788,46153846	25		
			Значение P	F критическое
			0,13120350077972	3,0491249886521

Рисунок 5.32. Результаты однофакторного ДА

Какой же из этого всего следует вывод? А вывод такой: поскольку вычисленное значение результата «F» (F-критерий) меньше, чем «F критическое» для данного уровня значимости, влияние фактора (технологии изготовления) на исследуемый параметр (долговечность лампочек) является несущественным.

- *Двухфакторный ДА*

В качестве примера применения Gnumeric для двухфакторного дисперсионного анализа рассмотрим задачу, приведенную в [Ю.А.Горицкий, Е.Е.Перцов, Практикум по статистике с пакетами StatGraphics, Statistica, SPSS. http://www.exponenta.ru/educat/systemat/goritskii/lr.asp].

Исследуется урожайность четырех сортов пшеницы, ц/га (фактор А, 4 уровня) от используемого вида удобрений (5 уровней фактора В). Данные получены с 20 участков равной площади и одинакового почвенного состава. Требуется выяснить влияние сорта пшеницы и типа удобрений на урожайность.

Таблицы исходных данных приведена на рис. 5.33.

	A1	A2	A3	A4
B1	19	25	17	21
B2	22	19	19	18
B3	26	23	22	25
B4	18	26	20	23
B5	21	22	21	24

Рисунок 5.33. Исходные данные для двухфакторного ДА

На рис. 5.34 показана вкладка «Ввод» диалога «Дисперсионный анализ: двухфакторный». Вкладки «Параметры» и «Вывод» являются стандартными, поэтому показывать их нет особого смысла.

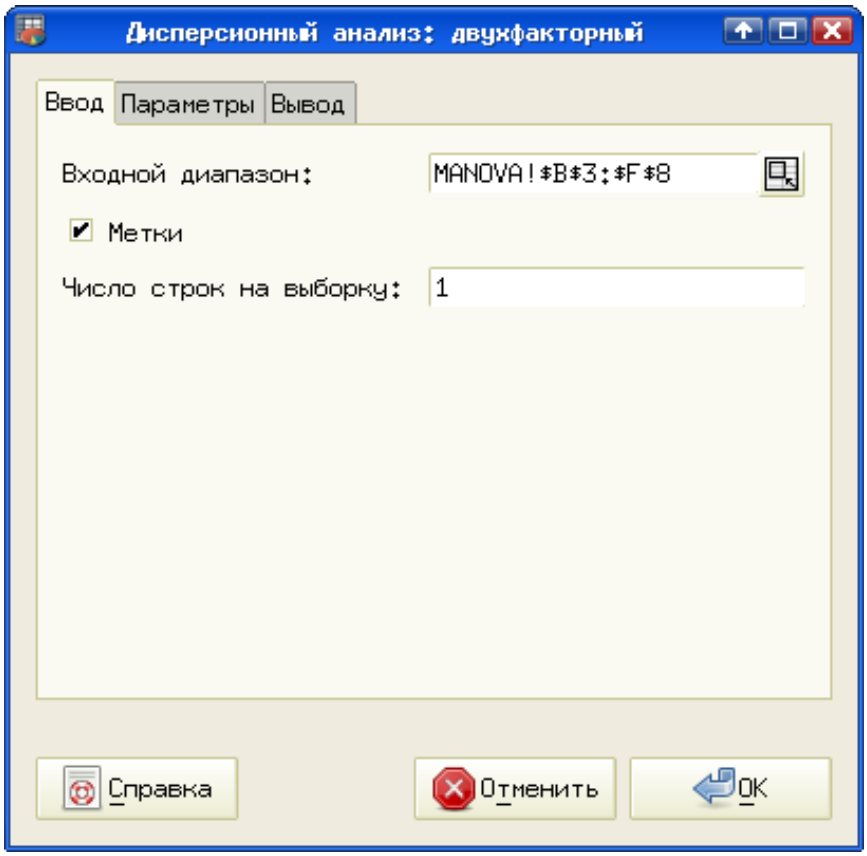

Рисунок 5.34. Определение исходных данных для двухфакторного ДА.

Результаты вычислений показаны на рис. 5.12. Для уменьшения размера рисунка итоговые значений («*F*», «*P*» и «*F критическое*») перенесены на другую строку.

Дисперсионный анализ: двухфакторный без воспроизведения				
Итог	Количество	Сумма	Среднее	Дисперсия
B1	4	82	20,5	11,6666667
B2	4	78	19,5	3
B3	4	96	24	3,33333333
B4	4	87	21,75	12,25
B5	4	88	22	2
A1	5	106	21,2	9,7
A2	5	115	23	7,5
A3	5	99	19,8	3,7
A4	5	111	22,2	7,7
Дисперсионный анализ				
Источник дисперсии	Сумма квадратов	степень свободы	Квадрат среднего	
Строки	46,2	4	11,55	
Столбцы	28,550000000000	3	9,51666666666679	
Ошибка	68,200000000000	12	5,6833333333333	
Всего	142,95	19		
	F	Значение P	F критическое	
	2,0322580645161	0,1536620984991	3,25916672690125	
	1,6744868035191	0,2251061113715	3,49029481949754	

Рисунок 5.35. Результаты двухфакторного ДА

При использовании Gnumeric получены те же значения уровней значимости *P* (0,153 и 0,225), что и в примере первоисточника. Соответственно, делается вывод о том, что дисперсионный анализ не обнаруживает влияния сорта пшеницы и типа удобрения на урожайность, что также видно из того, что при вычисленных уровнях значимости значения критерия F получаются меньше, чем соответствующие значения параметра «F критическое».

Нужно заметить, что использование свободно распространяемого пакета Gnumeric для решения подобных задач выглядит значительно привлекательнее использования пакета Statistica ценой около 700 USD, не говоря уже о StatGraphics или SPSS, легально приобрести которые весьма затруднительно.

5.11 Два средних

Вложенное меню «Два средних» («Сервис/Статистический анализ/Два средних») предоставляет набор инструментов для проверки гипотезы о равенстве (или неравенстве) средних значений выборок (генеральных совокупностей). В качестве исходных данных будем средствами Gnumeric генерировать выборки с нормальным распределением.

- *Равные выборки: Т-тест*

Пусть известно, что в двух выборках имеется равное количество значений случайных величин, и известно, что у этих выборок равные дисперсии. В данном случае Т-тест дает возможность определить дисперсии и средние значения для этих выборок и посчитать разницу между средними.

Для примера рассмотрим 25 нормально распределенных случайных значений со средним значением 5 и стандартным отклонением 1 (Выборка1) и 25 нормально распределенных случайных значений со средним значением 7 и стандартным отклонением 1 (Выборка2).

На вкладке «Ввод» диалога «Проверка различия двух средних» («Сервис/Статистический анализ/Два средних/Равные выборки: Т-тест...») указываем диапазоны для исходных данных, причем нужно проследить, чтобы все адреса были абсолютными (см. рис. 5.36). Режим «Метки», как всегда, позволяет использовать в выводе названия векторов данных.

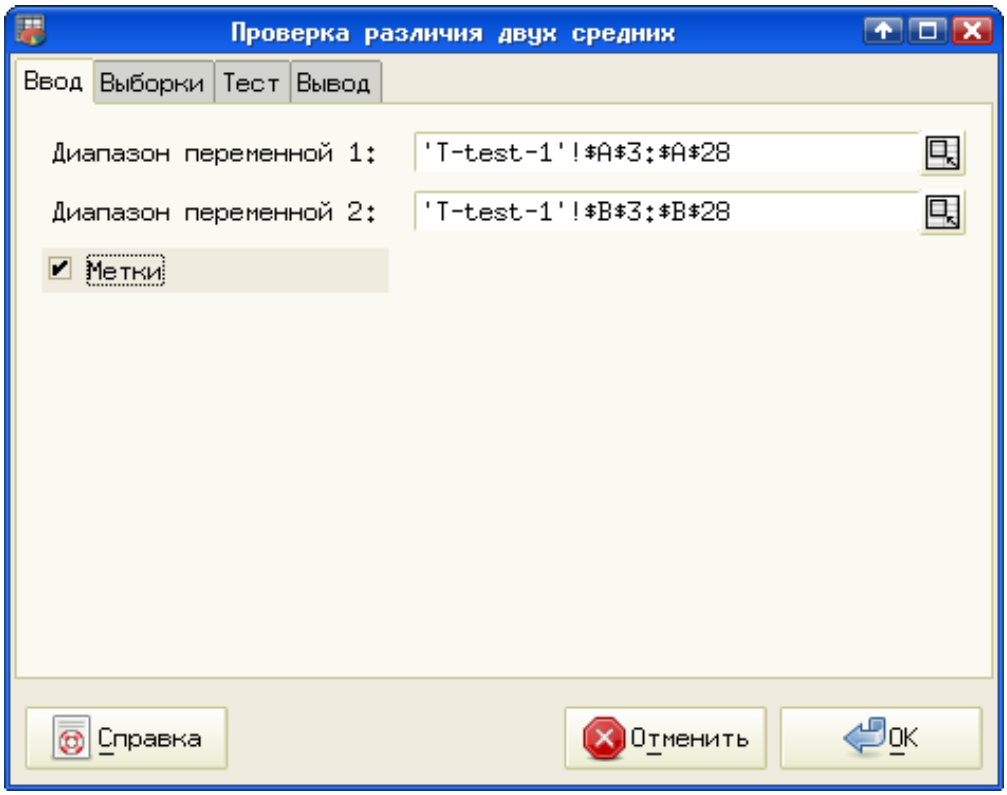

Рисунок 5.36. Настройка исходных данных для проверки гипотезы

На вкладке «Выборки» (рис. 5.37) указываем, что данные непарные, дисперсии неизвестны, но равны.

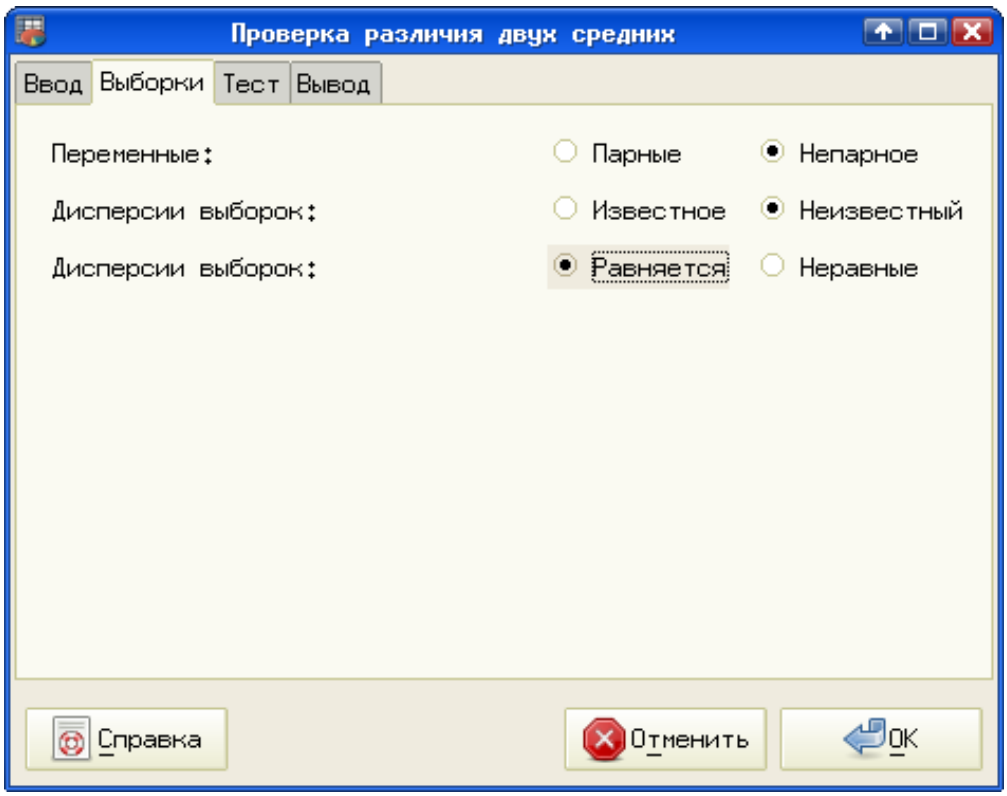

Рисунок 5.37. Настройка режимов анализа.

На рис. 5.38 показаны результаты вычислений, причем заметно прекрасное согласование результатов с заранее заданными параметрами выборок.

	Выборка1	*Выборка2*
Среднее	4,9856949420708	6,9219356509208
Дисперсия	1,00959256662225	1,1100760960711
Наблюдения	25	25
Накопленная дисперсия	1,05983433134665	
Гипотетическое среднее отклонение	0	
Наблюдаемое среднее отклонение	-1,93624070885001	
df	48	
t Stat	-6,64959754973821	
P (T<=t) одностороннее	1,2675474797465E-08	
t критическое одностороннее	1,67722419612551	
P (T<=t) двухстороннее	2,5350949594930E-08	
t критическое двухстороннее	2,01063475762452	

Рисунок 5.38. Результаты проверки гипотезы.

Проделав всё это с любыми модельными данными, в ячейках блока результатов можно увидеть формулы, по которым производятся расчеты.

- *Неравные выборки, равные дисперсии: Т-тест*

Теперь рассмотрим ситуацию, когда выборки имеют разное количество точек. Пусть параметры Выборки1 остаются прежними (25 точек, нормальное распределение, среднее значение 5 стандартное отклонение 1), а для Выборки2 установим следующие параметры – нормальное распределение со средним 7, стандартным отклонением 2 и количеством точек 20.

В этом случае на вкладке «Выборки» диалога «Проверка различия двух средних» все оставляем по умолчанию (рис. 5.39), и наблюдаем результаты (рис. 5.40).

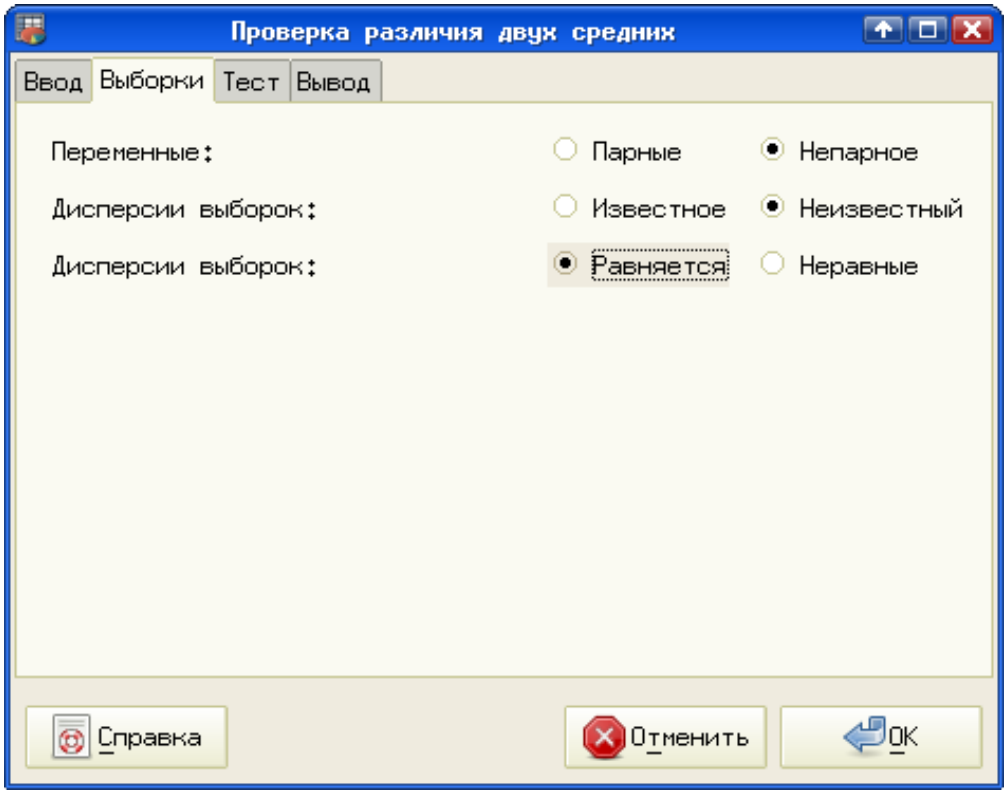

Рисунок 5.39. Настройка режимов вычислений.

Настройки режимов вычислений в этом случае совпадают с предыдущим случаем.

	Выборка1	Выборка2
Среднее	4,9290079315592	7,1298748506069
Дисперсия	0,88875584053372	0,9422275270696
Наблюдения	25	20
Накопленная дисперсия	0,91238286481703	
Гипотетическое среднее отклонение	0	
Наблюдаемое среднее отклонение	-2,20086691904767	
df	43	
t Stat	-7,68040227794376	
P (T<=t) одностороннее	6,752644595038E-10	
t критическое одностороннее	1,68107070320362	
P (T<=t) двухстороннее	1,350528919008E-09	
t критическое двухстороннее	2,01669219922857	

Рисунок 5.40. Результаты проверки гипотезы.

Опять-таки наблюдается трогательное соответствие результатов расчетов с заранее определенными параметрами выборок.

- *Неравные выборки, неравные дисперсии: T-тест*

Используя те же параметры распределения для Выборки1, что и в предыдущих случаях, для Выборки2 возьмем 20 точек, среднее значение 7 и стандартное отклонение 2.

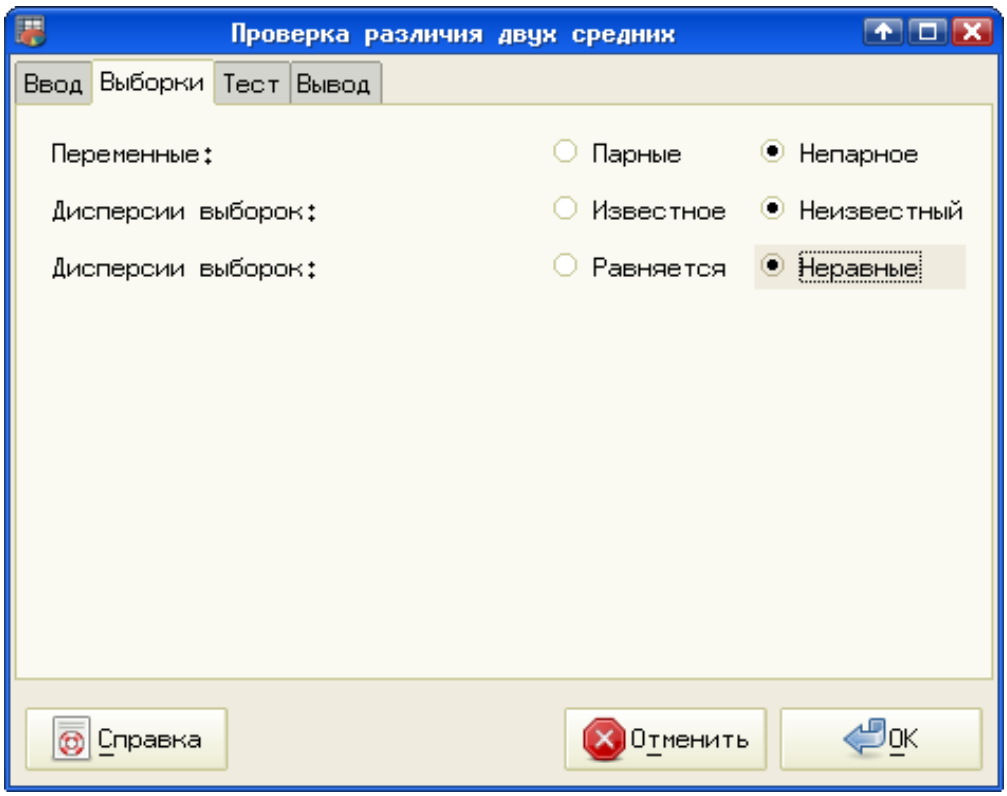

Рисунок 5.41. Настройка режимов вычислений.

На вкладке «Выборки» установим режимы в соответствии с рис. 5.41 и посмотрим результат (рис. 5.42).

	Выборка1	*Выборка2*
Среднее	5,22363347199752	6,9627418308079
Дисперсия	1,08564555028485	3,8771835295732
Наблюдения	25	20
Гипотетическое среднее отклонение	0	
Наблюдаемое среднее отклонение	-1,7391083588104	
df	27,3780733293963	
t Stat	-3,5701914147587	
P (T<=t) одностороннее	0,00067176549533	
t критическое одностороннее	1,7024535832589	
P (T<=t) двухстороннее	0,00134353099067	
t критическое двухстороннее	2,0505055360291	

Рисунок 5.42. Результаты проверки гипотезы.

- *Известные дисперсии: Z-тест*

Применим эту процедуру к слегка измененным модельным данным. Пусть Выборка1 остается с прежними параметрами, а для Выборки2 (среднее 7, 20 точек) установим дисперсию 2, для чего стандартное отклонение должно быть установлено

как 1,41.

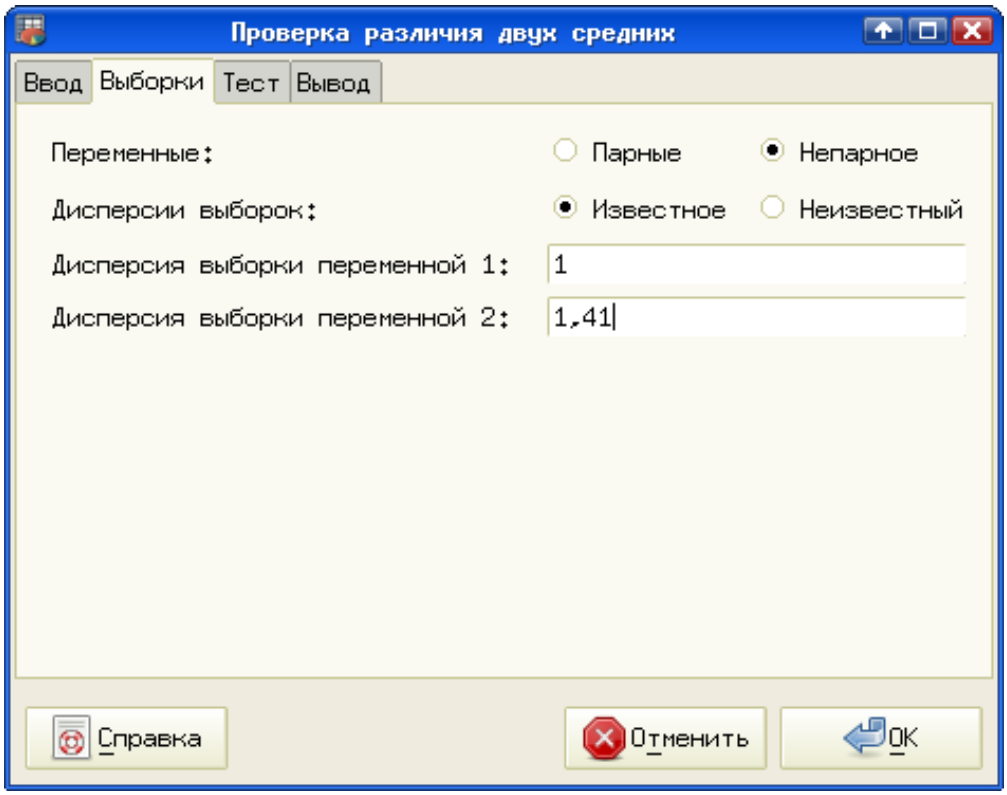

Рисунок 5.43. Настройка режимов вычислений.

На вкладке «Выборка» устанавливаем режимы и значения дисперсии в соответствии с рис. 5.43 и получаем результат, показанный на рис. 5.44.

	Выборка1	*Выборка2*
Среднее	4,81474951148839	6,523494868105
Известная дисперсия	1	2
Наблюдения	25	20
Гипотетическое среднее отклонение	0	
Наблюдаемое среднее отклонение	-1,70874535661631	
z	-4,56681406121364	
P (Z<=z) одностороннее	2,47596446779E-06	
z критическое одностороннее	1,64485362695147	
P (Z<=z) двухстороннее	4,95192893550E-06	
z критическое двухстороннее	1,95996398454005	

Рисунок 5.44. Результаты проверки гипотезы.

5.12 Две дисперсии: F-тест.

Этот инструмент позволяет проверить гипотезу о равенстве (или неравенстве) двух дисперсий. В качестве исходных данных будем использовать те же модельные

выборки, что и в случае проверки гипотезы о равенстве двух средних.

В качестве первого примера рассмотрим 25 нормально распределенных случайных значений со средним значением 5 и стандартным отклонением 1 (Выборка1) и 25 нормально распределенных случайных значений со средним значением 7 и стандартным отклонением 1 (Выборка2). В этом случае дисперсии однозначно равны.

Результат проведения теста показан на рис. 5.45.

F-Тест	Выборка1	Выборка2
Среднее	4,78087892427695	7,27042414268456
Дисперсия	1,37528345243102	1,44920309531595
Наблюдения	25	25
df	24	24
F	0,94899290297968	
P (F<=f) правостороннее	0,55049212345448	
F критическое правостороннее	1,98375956848961	
P (f<=F) левостороннее	0,44950787654552	
F критическое левостороннее	0,50409334673626	
P двухстороннее	0,89901575309104	
F критическое двухстороннее	0,44066893449361	2,26927727762143

Рисунок 5.45. F-тест. Равные дисперсии

Теперь рассмотрим вариант, при котором дисперсии отличаются в два раза (Выборка3 и Выборка4). Результат теста показан на рис. 5.46.

F-Тест	Выборка3	Выборка4
Среднее	5,1093984261493	4,62256516995347
Дисперсия	0,70692477100268	3,83214446689538
Наблюдения	25	25
df	24	24
F	0,18447236974221	
P (F<=f) правостороннее	0,99994957126822	
F критическое правостороннее	1,98375956848961	
P (f<=F) левостороннее	5,0428731778185E-05	
F критическое левостороннее	0,50409334673626	
P двухстороннее	0,00010085746356	
F критическое двухстороннее	0,44066893449361	2,26927727762143

Рисунок 5.46. F-тест. Неравные дисперсии

Вывод получается следующий: чем сильнее отличаются дисперсии выборок, тем меньше значение F.

5.13 Оценка выживаемости (оценка Каплана-Майера)

Общие сведения о задаче анализа выживаемости можно получить здесь [http://www.hr-portal.ru/statistica/gl14/gl14.php] или здесь [http://www.recognition.mccme.ru/pub/RecognitionLab.html/survival.pdf]. Суть задачи заключается в том, чтобы по набору признаков (характеристик) определить время сохранения объектом этих характеристик («время жизни») или распределение вероятностей сохранения характеристик в заданных пределах. Соответственно, можно строить прогнозы (предсказывать) среднее «время жизни» (время сохранения характеристик) таких объектов. Объектами могут быть вещества, устройства (приборы), сооружения и конструкции, а также живые существа. Чаще всего оценка выживаемости упоминается в связи с медицинской практикой.

В тех случаях, когда время наблюдения (продолжительность испытаний) меньше, чем «время жизни» конкретного объекта, получается, что «время жизни» точно не меньше времени наблюдения, а вот какое оно конкретно – узнать уже нельзя. Такие данные называются «цензурированными» (censored). Для группы объектов, участвующих в испытаниях возможны одновременно цензурированные и нецензурированные данные для различных экземпляров (например, при исследованиях срока службы энергосберегающих ламп в течение 10000 часов часть ламп вышла из строя в течение испытаний, а часть – так и не испортилась).

Пример использования Gnumeric для оценки выживаемости по Каплану-Майеру взят из справки по Gnumeric (Gnumeric 1.10.x).

Заготовим исходные данные в соответствии с рис. 5.47.

	A	B	C
1	Длительность	Группа	Цензурир.
2	1	1	1
3	1	2	1
4	2	1	1
5	2	2	1
6	3	2	0
7	3	1	1
8	4	1	1
9	4	1	0
10	4	2	1
11	4	2	1
12	5	1	0
13	6	1	0
14	7	1	1
15	9	2	0
16	11	2	1
17	12	1	0
18	12	2	0
19	13	2	1
20	18	1	0
21	19	2	0

Рисунок 5.47. Пример исходных данных для анализа выживаемости

Первый столбец («Длительность») означает время испытаний (наблюдений) для каждого исследуемого экземпляра. В столбце «Группа» задаётся принадлежность объекта к группе объектов (группы могут отличаться местоположением, периодом времени наблюдений и другими признаками и обстоятельствами). В данном примере имеется только две группы. Наконец, в третьем столбце указывается признак «цензурированности» данных (если в ячейке 1 – данные цензурированы).

Все данные носят дискретный характер («время жизни» изменяется дискретно).

Диалог настройки анализа вызывается через вложенное меню «Данные/Статистический анализ». Сначала определяется набор исходных данных и их цензурированность (вкладка «Ввод» диалога, рис. 5.48). Использование цензурированных данных разрешается включением соответствующего режима (Permit censorship).

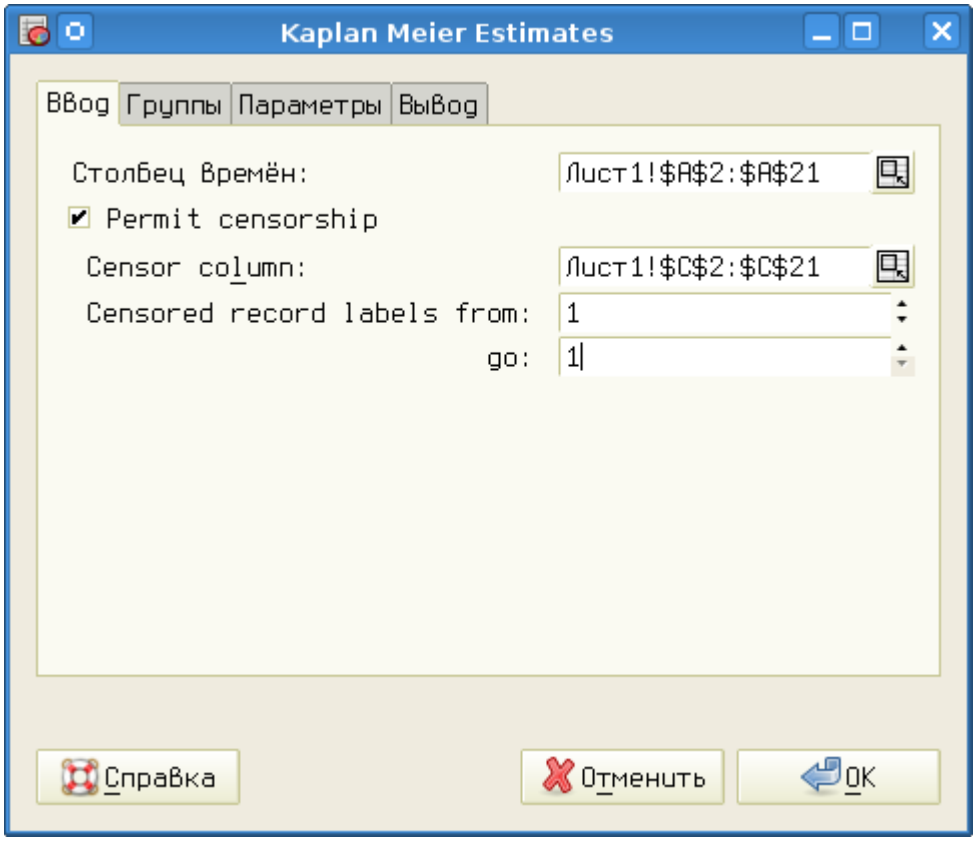

Рисунок 5.48. Настройка исходных данных для анализа

На вкладке «Группы» задаётся количество групп и номера, которые их определяют. Теоретически можно объединять несколько групп в одну, указав диапазон номеров «от» и «до» (рис. 5.49). Для установки номера группы используются поля со счётчиками.

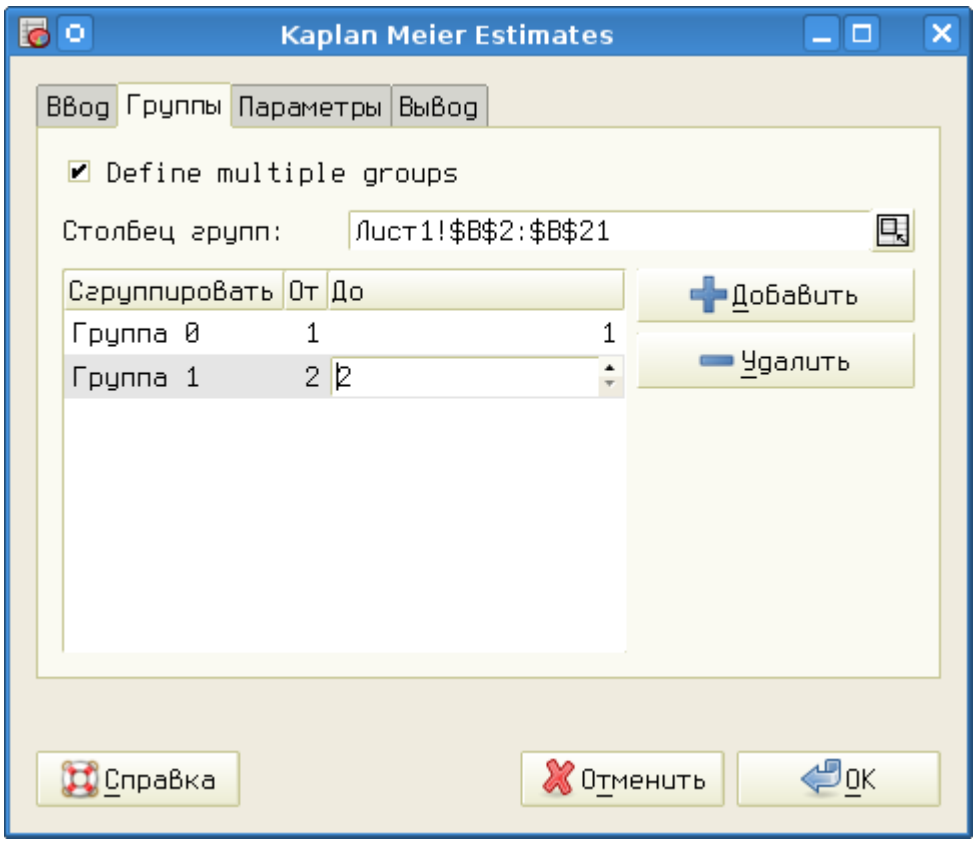

Рисунок 5.49. Настройка групп исследуемых объектов

В этом примере (и по умолчанию) используется две группы, но с помощью кнопок «+Добавить» и «-Удалить» количество групп можно изменять так, как требуется.

На вкладке «Параметры» (рис. 5.50) определяется объём итоговой информации. Различные виды результатов можно включать и выключать. Пусть в рассматриваемом примере будет выводиться максимально полный набор результатов.

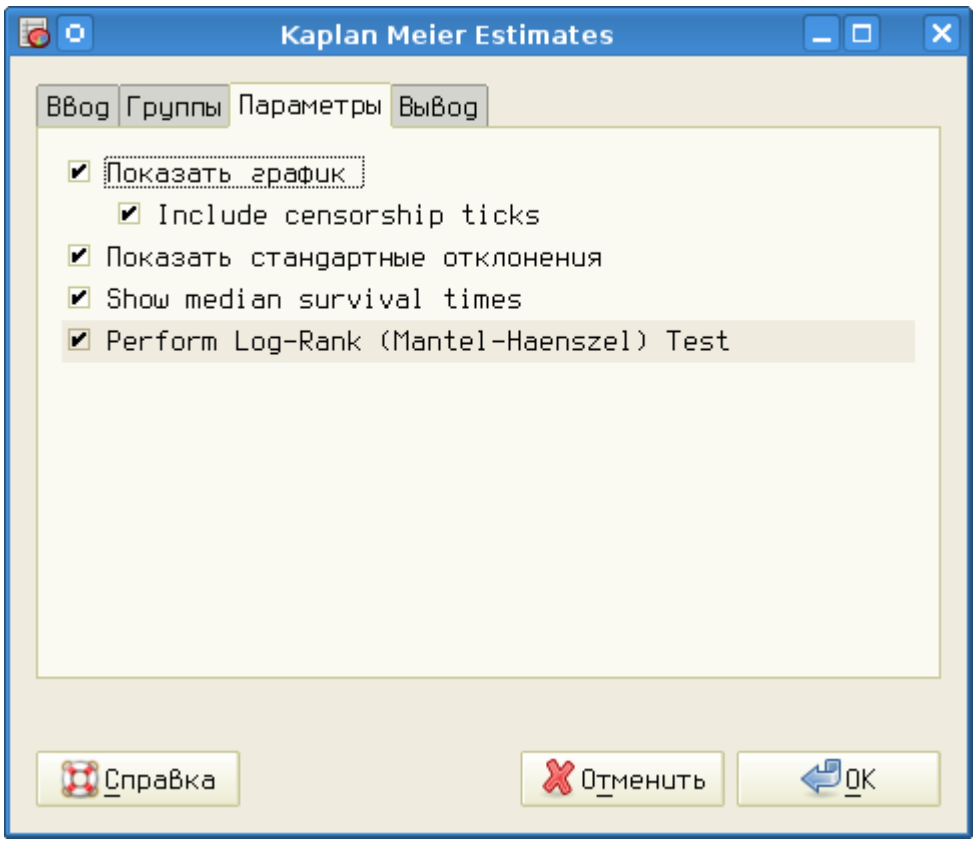

Рисунок 5.50. Настройка результатов анализа

Наконец, на вкладке «Вывод» (рис. ?) имеет смысл выбрать вариант создания нового листа, поскольку количество выводимых результатов достатоно велико.

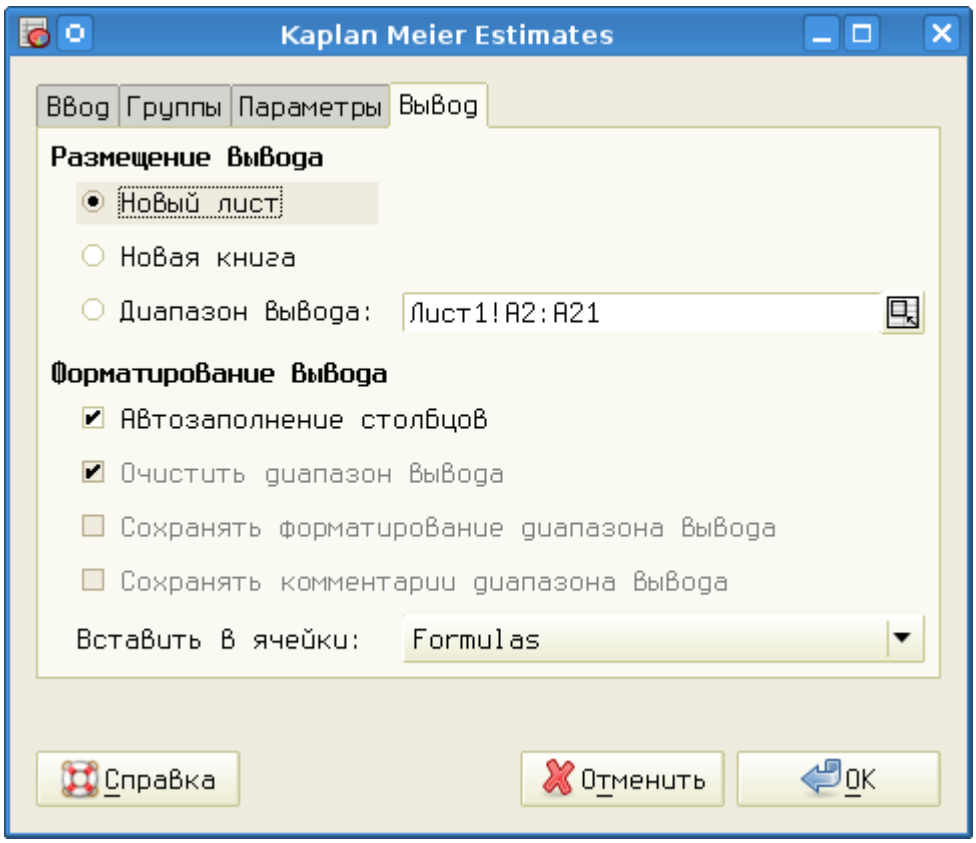

Рисунок 5.51. Определение расположения результатов анализа

В результате получается график, на котором отмечены точки с цензурированными данными для обоих групп (рис. 5.52), а также выдаются численные результаты. На рис. 5.52 результаты для первой группы показаны сплошной линией, цензурированные точки – треугольниками, а результаты для второй группы – «точечной» линией, цензурированные точки – ромбы.

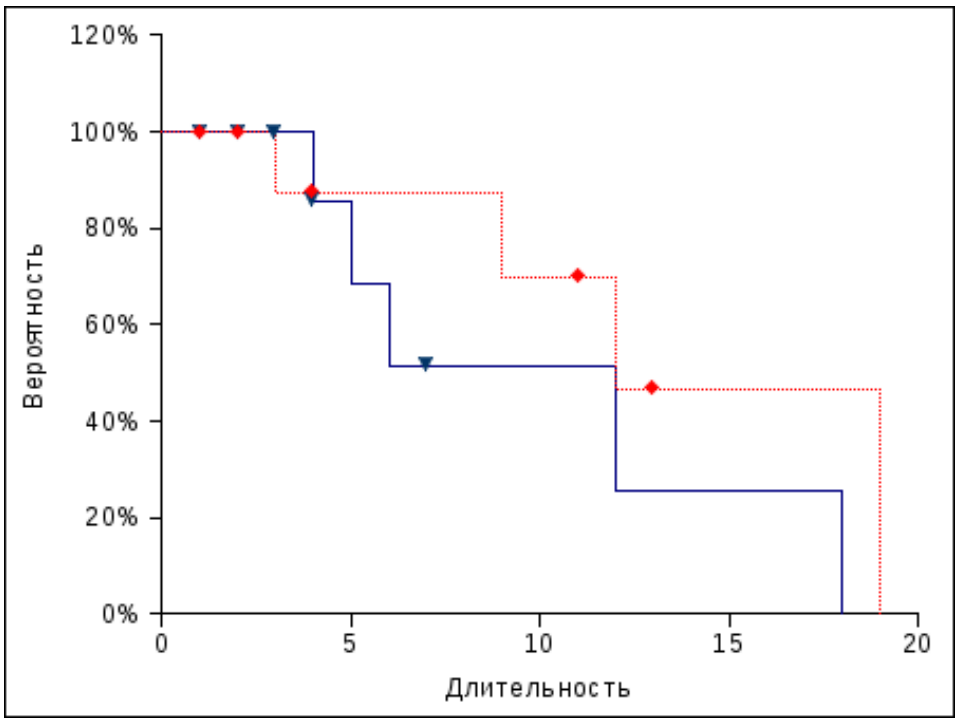

Рисунок 5.52. Результаты анализа (функция выживания)

Численные результаты для первой группы показаны на рис. 5.53. Наличие деления на 0 при времени в 19 единиц, видимо, связано с тем, что для первой группы («Группа0») нет нецензурированных данных для такого «време жизни».

	A	B	C	D	E	F
1	**Каплан-Мейер**	*Группа 0*				
2	*Время*	*At Risk*	*Deaths*	*Censures*	*Probability*	*Стандартное отклонение*
3	0	10	0	0	100,00%	0,0000
4	1	10	0	1	100,00%	0,0000
5	2	9	0	1	100,00%	0,0000
6	3	8	0	1	100,00%	0,0000
7	4	7	1	1	85,71%	0,1224
8	5	5	1	0	68,57%	0,1719
9	6	4	1	0	51,43%	0,1792
10	7	3	0	1	51,43%	0,2069
11	9	2	0	0	51,43%	0,2534
12	11	2	0	0	51,43%	0,2534
13	12	2	1	0	25,71%	0,1567
14	13	1	0	0	25,71%	0,2216
15	18	1	1	0	0,00%	0,0000
16	19	0	0	0	#Деление на 0!	#Деление на 0!

Рисунок 5.53. Численные результаты для первой группы

В следующих столбцах располагаются результаты для второй группы («Группа1»). Для получения иллюстрации столбцы таблицы от B до F были скрыты (рис. 5.54).

	A	G	H	I	J	K
1	**Каплан-Мейер**	Группа 1				
2	Время	At Risk	Deaths	Censures	Probability	Стандартное отклонение
3	0	10	0	0	100,00%	0,0000
4	1	10	0	1	100,00%	0,0000
5	2	9	0	1	100,00%	0,0000
6	3	8	1	0	87,50%	0,1094
7	4	7	0	2	87,50%	0,1169
8	5	5	0	0	87,50%	0,1383
9	6	5	0	0	87,50%	0,1383
10	7	5	0	0	87,50%	0,1383
11	9	5	1	0	70,00%	0,1715
12	11	4	0	1	70,00%	0,1917
13	12	3	1	0	46,67%	0,1968
14	13	2	0	1	46,67%	0,2410
15	18	1	0	0	46,67%	0,3408
16	19	1	1	0	0,00%	0,0000

Рисунок 5.54. Численные результаты для второй группы

Наконец, общее сравнение среднего времени выживаемости в группах обеспечивается тестом Log-Rank (рис. 5.55).

	M	N	O
		Группа 0	Группа 1
	Медиана:	12	12
	Log-Rank Test:		
	Statistics:	1,00696511856098	
	Degrees of Freedom:	1	
	p-Value:	0,31563100206861	

Рисунок 5.55. Общая статистика по группам

Значение p позволяет оценить различие среднего времени жизни по группам. На основании полученной в рассматриваемом примере величины p делается вывод, что эти значения статистически неразличимы.