**Лабораторная работа № 1**

**«ЗНАКОСТВО С ПРОГРАММЫМ ПАКЕТОМ STATISTICA»**

*Создание файла данных в STATISTICA*

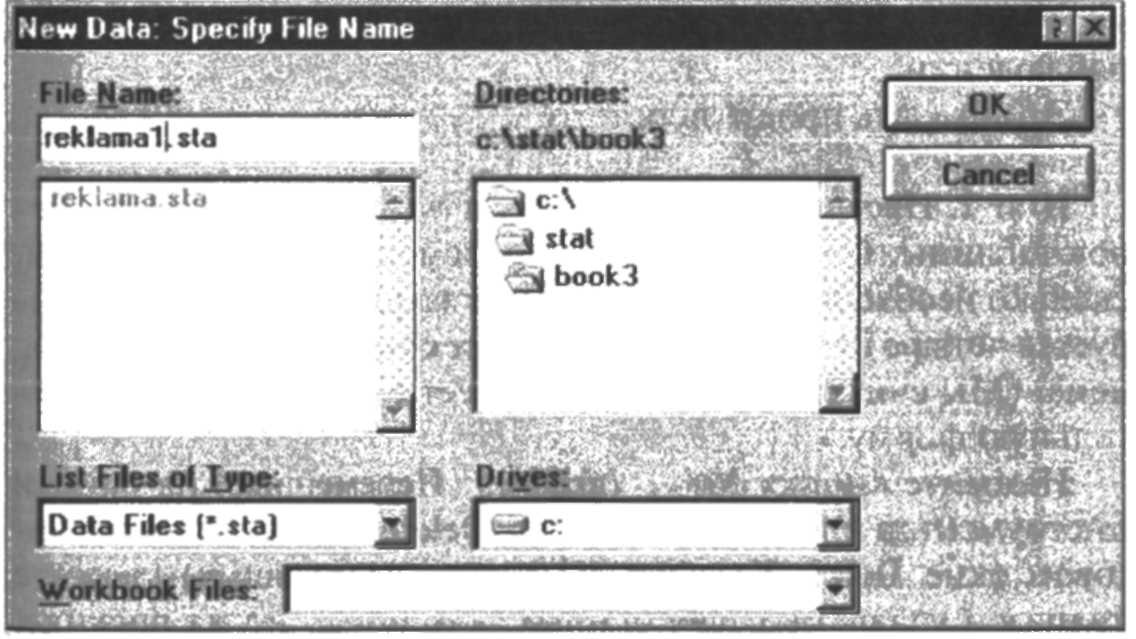
Задание 1. В следующей таблице приведены данные о размере рекламного объявления и его цене при размещении на полосе газеты:

|  |  |  |  |
| --- | --- | --- | --- |
| ДЛИНА (мм) | ШИРИНА (мм) | ПЛОЩАДЬ | ЦЕНА (руб.) |
| 47 | 35 | 1645 | 1446000 |
| 47 | 73 | 3431 | 2768000 |
| 47 | 111 | 5217 | 3974000 |
| 47 | 149 | 7003 | 5147000 |
| 47 | 209 | 9823 | 6290000 |
| 47 | 225 | 10575 | 7537000 |
| 47 | 263 | 12361 | 8828000 |
| 47 | 301 | 14147 | 10260000 |

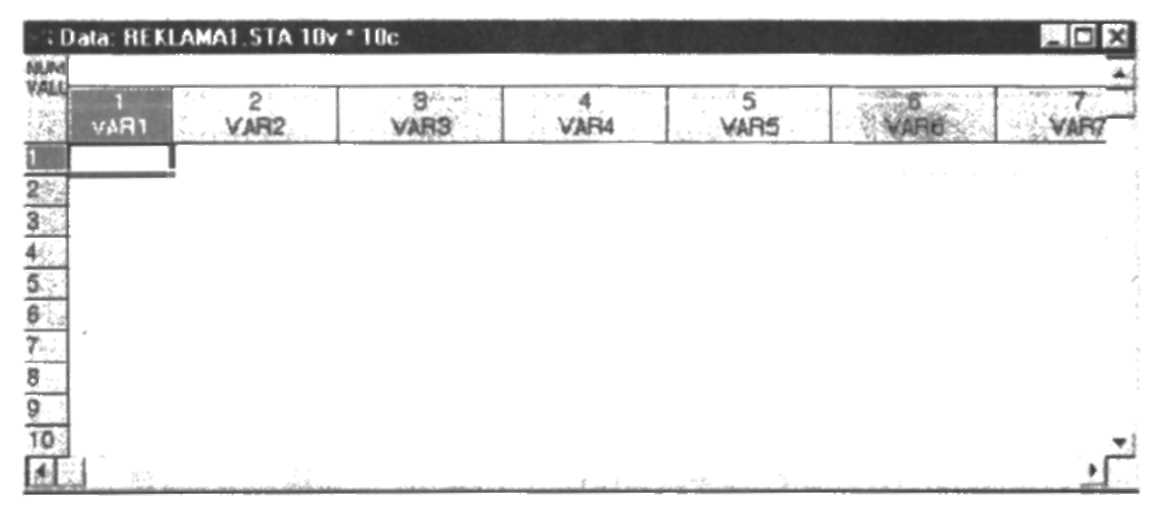
Создайте файл STATISTICA с этими данными.

*Решение:*

**Шаг 1.** Создание электронной таблицы. Выберите команду **New Data (Новые данные)** из меню **File (Файл).** Эта команда доступна также по комбинации клавиш **CTRL+N.** В появившемся диалоговом окне **New Data: Specify File Name (Новые данные: Определить имя файла)** введите имя: reklama1.sta.



Нажмите кнопку **OK**. STATISTICA автоматически откроет пустую электронную таблицу с именем reklama1.sta, которая и появится на экране.

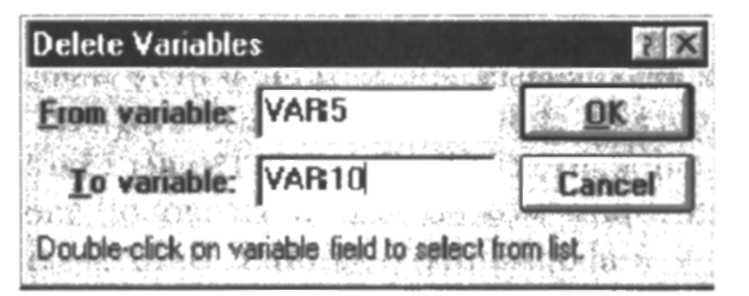


В заголовке окна электронной таблицы автоматически отображается имя файла и его размер (reklama1.sta 10v\*10c). Размер таблицы по умолчанию принят 10 на 10 (10 переменных с именами VAR1, VAR2, VAR3... VAR10 и10 пронумерованных наблюдений).

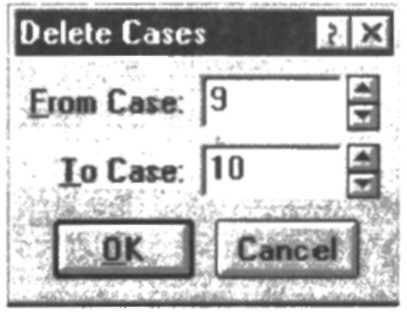
Сделайте в таблице столько строк и столбцов, сколько нужно. Нам нужно, чтобы в таблице имелось 4 переменные и 8 наблюдений (согласно таблице).

**Шаг 2.** Настройка таблицы. Произведем настройку размеров таблицы. Создадим столько переменных и наблюдений, сколько необходимо. Для данных рекламного объявления требуется четыре переменных: ширина рекламы, длина, площадь, цена.

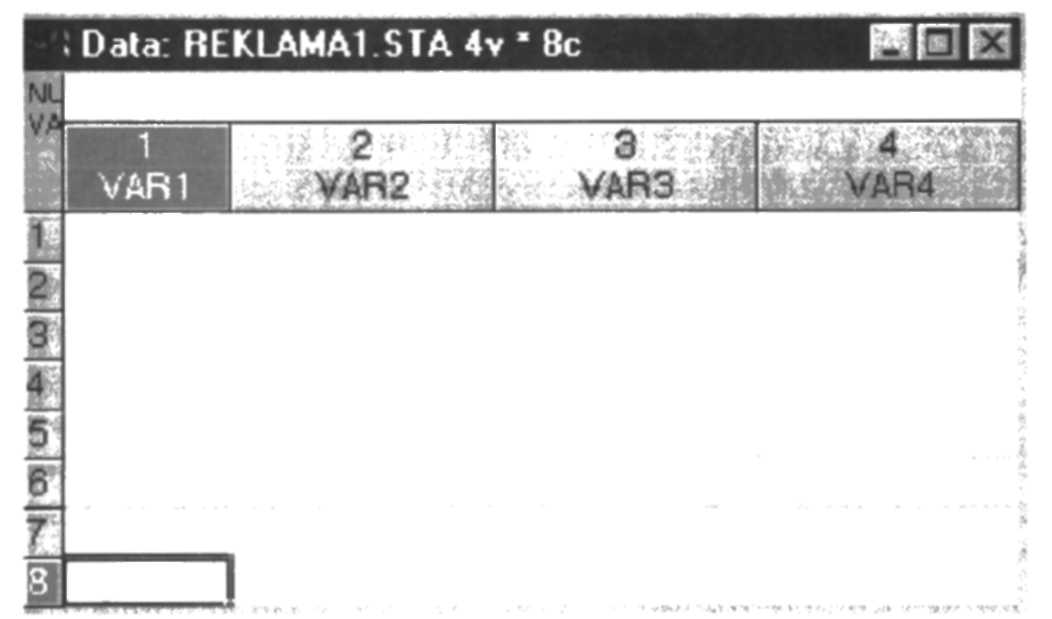
Нажмите кнопку **[Var]** **Variables (Переменные)** на панели инструментов и выберите команду **Delete (Удалить)**. В диалоговом окне **Delete Variables (Удаление переменных)** укажите диапазон удаляемых переменных, как показано на рисунке. Нажмите кнопку **ОК**.



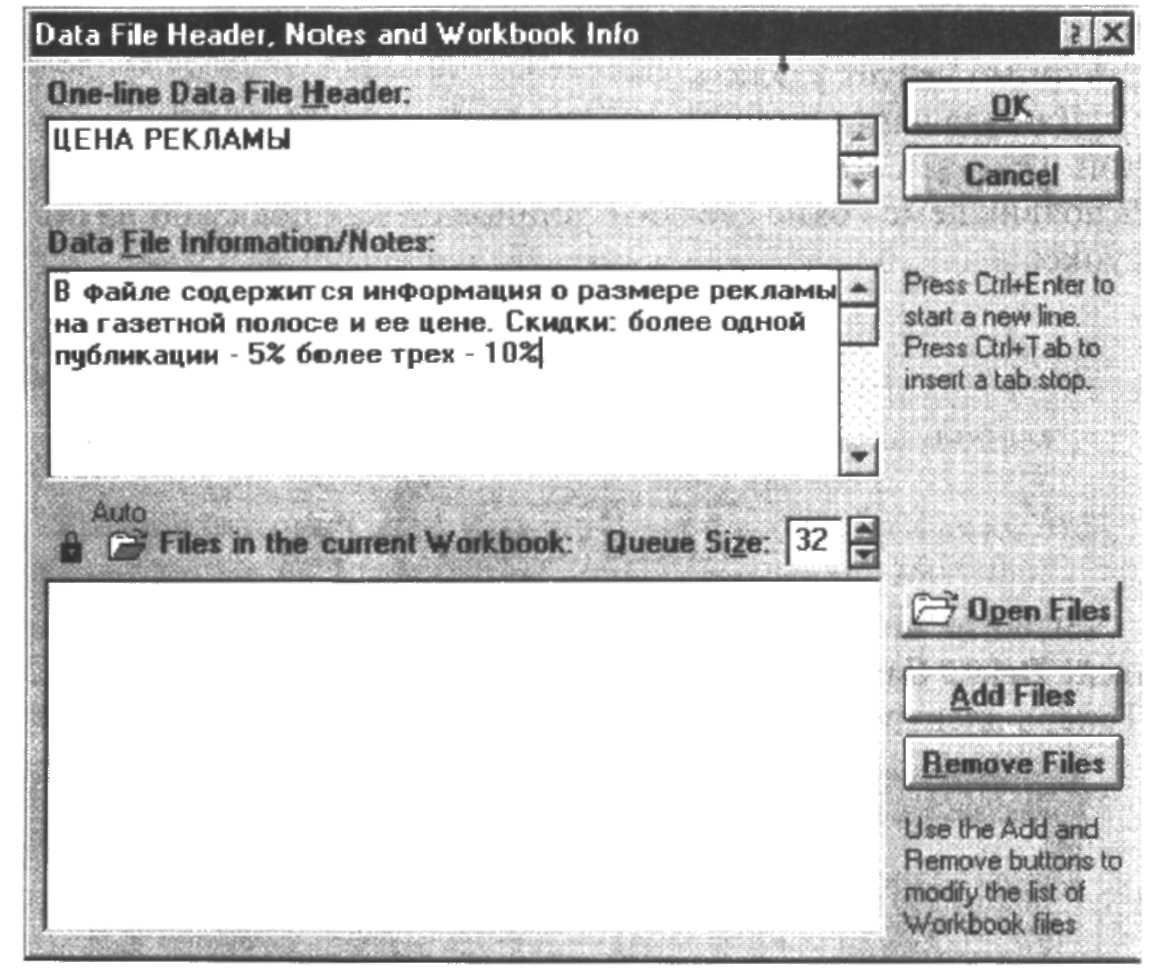
Необходимое число наблюдений – 8. В созданной таблице число наблюдений равно 10. Два лишних наблюдения из таблицы следует удалить. Для удаления лишних наблюдений воспользуйтесь кнопкой **[Cases]** **Cases (Наблюдения)** и командой **Delete (Удалить)**. В появившемся окне сделайте установки, как показано на рисунке:



Задайте диапазон удаляемых наблюдений в диалоговом окне **Delete Cases (Удалить наблюдения)**. Нажмите кнопку **ОК**. Теперь электронная таблица выглядит следующим образом



**Шаг 3.** Подготовка таблицы к вводу данных, заголовок таблицы и имена переменных. Дважды щелкните мышью на белом поле в таблице под словами: Data: REKLAMA1.STA 4V\*8C. На экране появится окно **Data File Header (Заголовок файла данных)**, в котором можно задать заголовок таблицы и дополнительную информацию о данных. Введем заголовок таблицы: ЦЕНА РЕКЛАМЫ и дополнительную информацию о скидках на рекламу, как показано на рисунке ниже:



Эта дополнительная информация всегда будет доступна вам.

Нажмите кнопку **ОК** в диалоговом окне **Data File Header (Заголовок файла данных)**. Теперь таблица выглядит следующим образом:

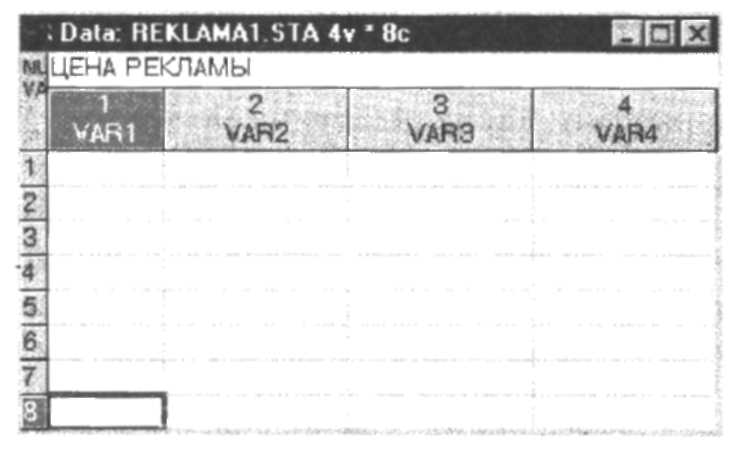
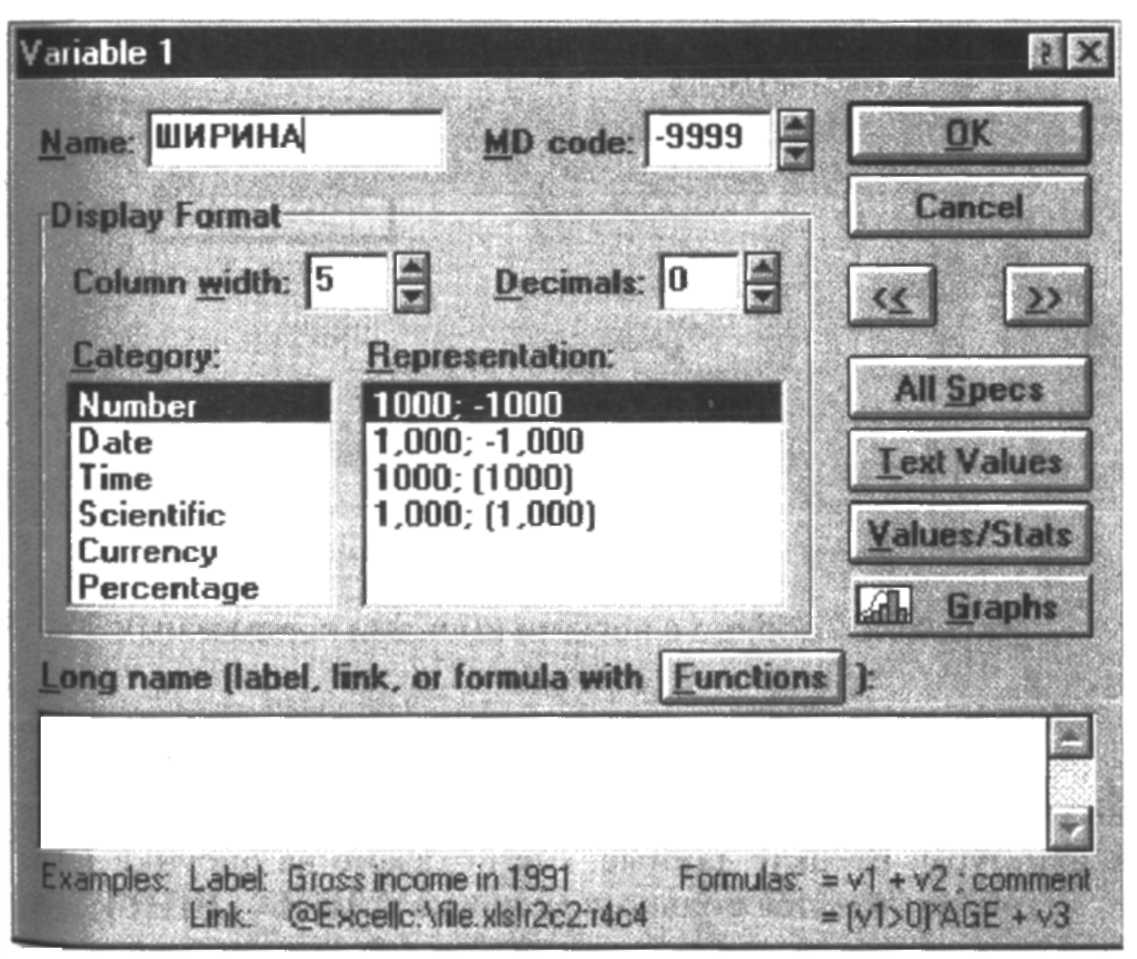
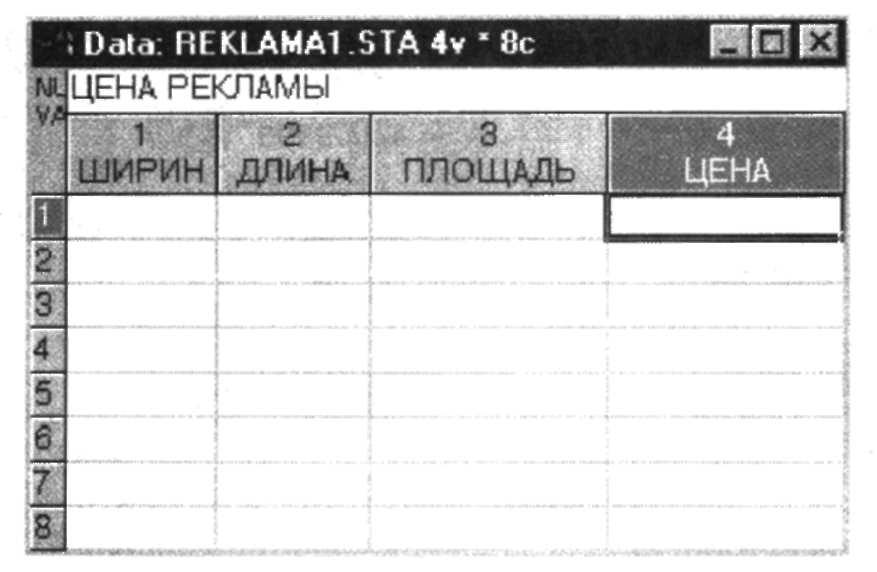


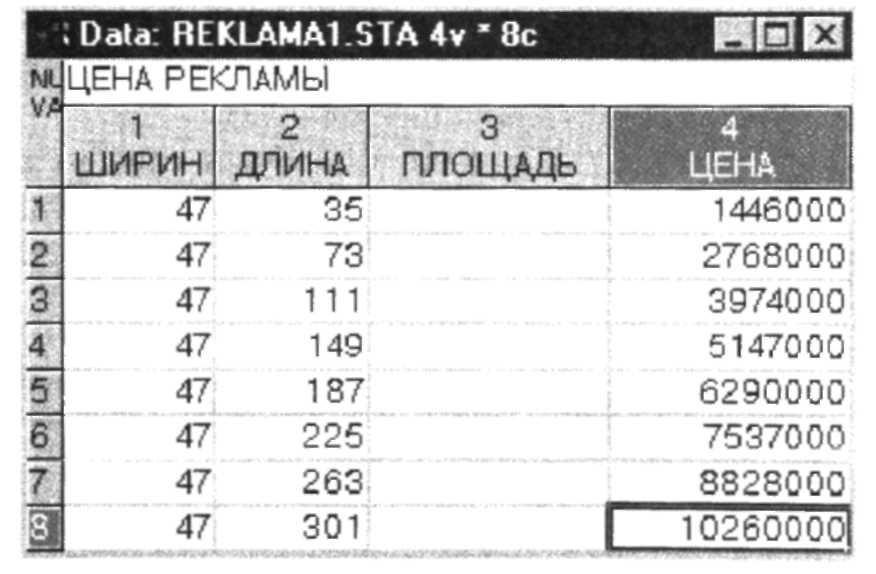
Таблица почти готова к вводу данных, однако придадим ей еще несколько более удобный вид: введем имена переменных, которые отражают смысл записей, и специфицируем их. Дважды щёлкните на имени переменной **VAR1** в электронной таблице. На экране появится окно спецификации переменной VAR1. В поле **Name (Имя)** напишите: ШИРИНА. Длина имени не должна превышать 8 символов.



Нажмите кнопку **ОК**. То же сделайте для переменной VAR2, ей присвойте имя ДЛИНА, переменной VAR3 присвойте имя ПЛОЩАДЬ. Переменной VAR4 присвойте имя ЦЕНА. Для двух последних переменных в рамке Display Format (Формат отображения) задайте Desimals (Число разрядов после десятичной точки) равным нулю, а в рамке Column width (Ширина столбца) – 8. Теперь таблица готова к тому, чтобы ввести в нее данные.

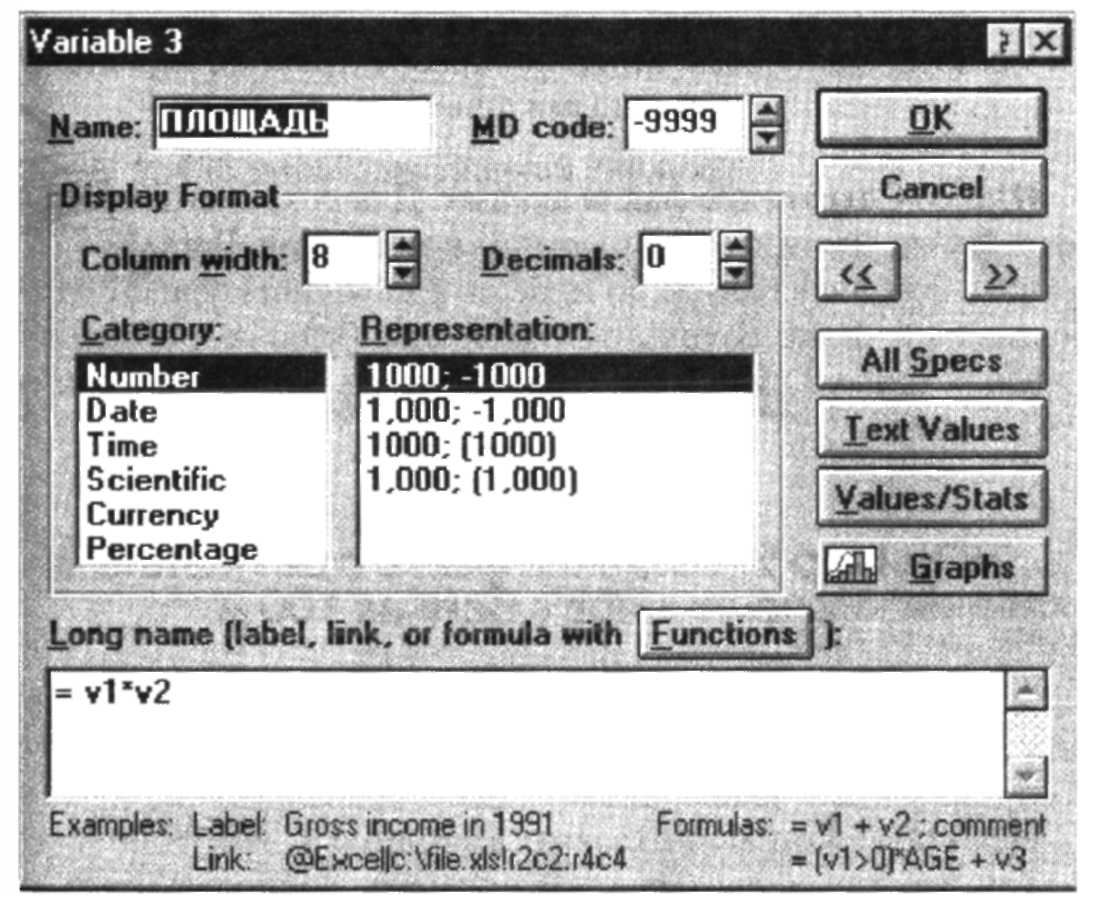


**Шаг 4.** Ввод данных в электронную таблицу. Так как данных немного, будем вводить их с клавиатуры. Введите данные, как показано на рисунке:

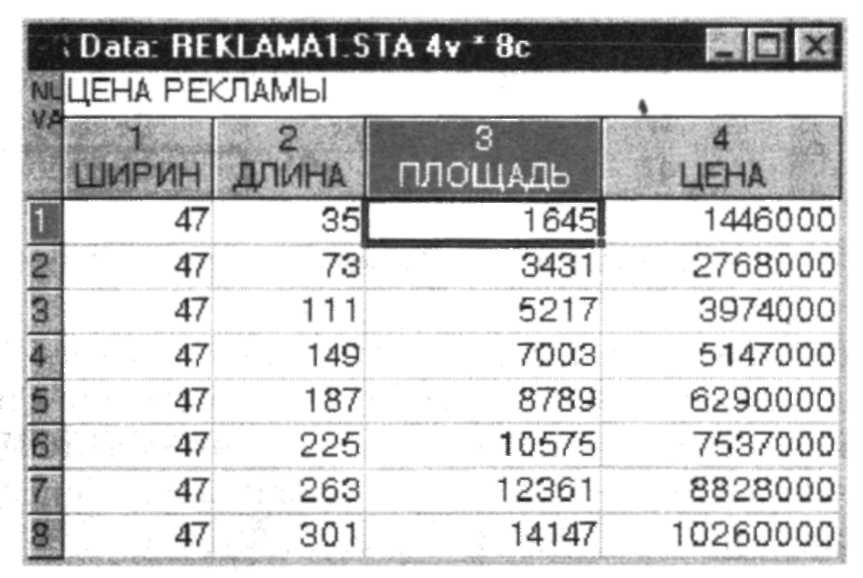


Данные вводятся только в колонки ДЛИНА, ШИРИНА, ЦЕНА. Колонка ПЛОЩАДЬ остается пустой. Введем данные в колонку ПЛОЩАДЬ. Сделаем это прямым подсчетом площади прямоугольника по длине и ширине.

Дважды щелкните на имени переменной ПЛОЩАДЬ в электронной таблице. На экране появится окно спецификации переменной ПЛОЩАДЬ.



В поле **Long names (Длинные имена)** запишите формулу **=v1\*v2**. Нажмите кнопку **ОК**. Площадь рекламных объявлений будет подсчитана и занесена в ячейки переменной VAR3. Полностью заполненная таблица стоимости рекламных объявлений на полосе газет появится на вашем экране:



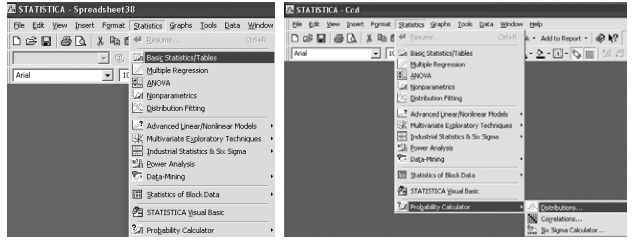
**Шаг 5**. Сохранение файла данных. Для сохранения созданного файла нажмите мышью на кнопку **Save Data File (Сохранить файл данных)** либо наберите на клавиатуре **CTRL+S**.

Задание 2. Создание таблицу с результатами олимпийских чемпионов в беге на 100 м:

|  |  |  |  |
| --- | --- | --- | --- |
| Год | Чемпион | Страна | Время |
| 1896 | Вэрк | США | 12.0 |
| 1900 | Джервис | США | 10.8 |
| 1904 | Хан | США | 11.0 |
| 1906 | Хан | США | 11.2 |
| 1908 | Уолкер | ЮАР | 10.8 |
| 1912 | Крейг | США | 10.8 |
| 1920 | Пэддок | США | 10.8 |
| 1924 | Абрахаме | Англия | 10.6 |
| 1928 | Уияльямс | Канада | 10.8 |
| 1932 | Тоулэн | США | 10.3 |
| 1936 | Оуэнс | США | 10.3 |
| 1948 | Диллард | США | 10.3| |
| 1952 | Реминджил | США | 10.4 |
| 1956 | Морроу | США | 10,5 |
| 1960 | Хари | ФРГ | 10.2 |
| 1964 | Хейес | США | 10.0 |
| 1968 | Хайнс | США | 9.9 |
| 1972 | Борзов | СССР | 10,1 |
| 1976 | Кроуфорд | Тринидад | 10.6 |

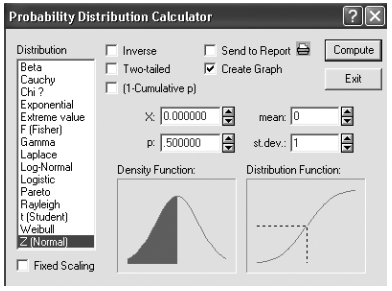
*Вероятностный калькулятор*

Запустите модуль **Basic Statistics/Tables (Основные статистики/таблицы)** из переключателя модулей. Высветите в стартовой панели модуля **Basic Statistics/Tables (основные статистики/таблицы)** строку **Probability calculator (вероятностный калькулятор)** (рис. 1, 2).



*Рис. 1 Рис. 2*

Нажмите кнопку **ОК**. Откроется окно **Probability Distribution Calculator (калькулятор вероятностных распределений)** (рис. 3).

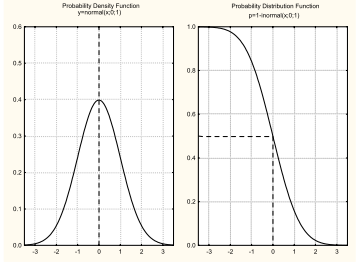


*Рис. 3 – Калькулятор вероятностных распределений*

Окно имеет следующую структуру – в левой части список распределений **Distribution (распределение)**. Многие стандартные распределения в этом окне можно выбрать, высвечивая их названия в списке слева: *Бета*, *Коши*, *хи-квадрат*, *нормальное*, *логнормальное,* *распределение Стьюдента* и т. д. Выберем в списке строчку *Z(Normal) – нормальное распределение*. Автоматически справа появляются поля, где можно задать параметры нормального распределения: **mean (среднее)** и **st. dev (стандартное отклонение)** (рис. 3). Система по умолчанию запишет в них стандартные значения: **среднее = 0, стандартное отклонение = 1**. Эти значения можно изменить. Одновременно с выбором распределения в левом списке справа в калькуляторе появляются графики нормальной плотности и функции распределения: **Density Function (функция плотности), Distribution Function (функция распределения)**.

В поле **р** задается уровень вероятности. Поместите курсор мыши в это поле и щелкните левой кнопкой. Наберите любое значение в интервале от 0 до 1. После нажатия на кнопку **Compute (вычислить)** в строке **X** появится соответствующий квантиль. То же можно сделать и в обратную сторону.

Опции в верхней части окна имеют следующее назначение: **Inverse (обратная функция распределения)**, **Two-tailed (двухсторонний)**, **1-Cumulative р**, **Send to Print (отправить на печать)**, **Create graph (создать график)**. Если пометить опцию **Create graph (создать график)** и нажать далее кнопку **Compute (вычислить)**, то на экране появится график плотности и функции распределения (задайте в строке р значение 0) (рисунок 4).



*Рисунок 4 – Плотность и функция распределения стандартной нормальной величины*

Вероятностный калькулятор заменяет многие таблицы. Вместо того чтобы использовать таблицы распределений можно использовать данный калькулятор.

*Нормальное распределение*

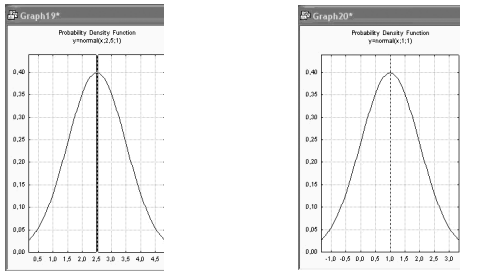
Наиболее часто встречающееся в статистике распределение – это нормальное распределение. Нормальное распределение имеет два параметра: **mean – среднее**; **standard deviation – стандартное отклонение**. Эти параметры задаются в окне вероятностного калькулятора.

Плотность нормального распределения симметрична относительно среднего. Среднее значение определяет меру расположения плотности, и при нормальном распределении совпадает с медианой и модой.

Зададим различные значения среднего, оставив пока без изменения стандартное отклонение. Будем считать, что оно равно 1.

**Шаг 1.** Откройте вероятностный калькулятор в поле **mean (среднее)**, задав вначале 1. В поле **р** задайте значение 0,5 (в данном примере это чисто техническая установка).

**Шаг 2.** Выберите опцию **Create graph (создать график)** и нажмите кнопку **Compute (вычислить)**, на экране появится график плотности (рис. 5а).

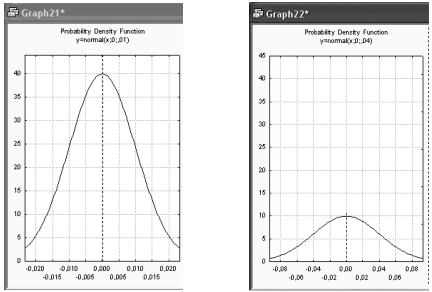


а) б)

*Рисунок 5.5 – Плотность нормального распределения со средним а) 1; б) 2,5*

Повторите те же действия, задав в поле mean (среднее) значение 2,5. Вы увидите следующий график (рис. 5б). Обратите внимание, что график плотности нормального распределения сдвигается по оси ординат при изменении среднего – при возрастании среднего графики сдвигаются вправо.

Посмотрите, как меняется плотность распределения при изменении другого параметра – стандартного отклонения. Для этого задайте различные значения стандартного отклонения, считая, что среднее фиксировано и равно 0 (рис. 6 а, 6 б).



а) б)

*Рис.6 – Плотность нормального распределения со средним 0 и дисперсией а) 0,01 и б) 0,04*

При увеличении дисперсии плотность нормального распределения расплывается или рассеивается относительно среднего значения, при уменьшении дисперсии она наоборот сжимается, концентрируясь возле одной точки – точки максимального значения.

*Рассмотрим пример использования нормального распределения*

***Задача.*** В некоторой стране рост взрослых мужчин приближенно имеет нормальное распределение со средним 176,6 см и стандартным отклонением 7,63 см. Какова вероятность того, что рост наугад выбранного мужчины не больше 185 см и не меньше 175 см?

Решение:

**Шаг 1.** Откройте вероятностный калькулятор. Выберите в списке распределений **Z(Normal) (нормальное распределение).**

**Шаг 2.** Задайте в поле **mean – среднее** 175,6; в поле **st.dev. – стандартное отклонение** 7,63.

**Шаг 3.** В поле **Х** задайте 185. Нажмите кнопку **Compute (вычислить).** В поле **р** появилось значение 0,891022. Запомните это значение как p1.

**Шаг 4.** В поле **Х** задайте 175. Нажмите кнопку **Compute (вычислить)**. В поле **р** появилось значение 0,468661. Запомните это значение как р2.

**Шаг 5.** Вычтите р2 из p1. Вы получите 0,422361.

Итак, с вероятностью 0,422361 встреченный вами мужчина имеет рост не ниже 175 и не выше 185 сантиметров.

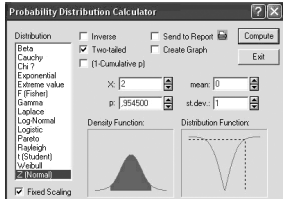
*Правила 2 и 3 сигма*

Правила 2 и 3 сигма необходимо знать (они часто используются на практике). Смысл состоит в том, что если от точки среднего, или от точки максимума плотности нормального распределения отложить вправо и влево соответственно 2 и 3 стандартных отклонения (2 и 3 σ), то площадь под графиком нормальной плотности, подсчитанная по этому промежутку, будет соответственно равна 95,45% и 99,73% всей площади под графиком. То есть 95,45% и 99,73% всех независимых наблюдений из нормального распределения лежат в пределах 2 и 3 стандартных отклонений от среднего значения.

Проверьте это правило с помощью вероятностного калькулятора.

**Шаг 1.** Выберите нормальное распределение в списке распределений

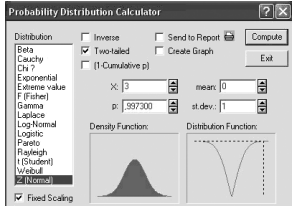
**Шаг 2.** Задайте стандартные параметры: среднее 0, стандартное отклонение 1, пометьте **опцию Two-tailed (двухсторонний)**, в строке **X** задайте 2 (2 стандартных отклонения), нажмите **Compute**, в строке **р** появится значение 0,954500 (рис. 7).



*Рис. 7 – Иллюстрация к правилу 2 сигм*

В поле **Density Function (функция плотности)** вероятностного калькулятора показана заштрихованная площадь под графиком плотности, в поле **р** показано значение 0,9545 (95,45%). Заштрихованная площадь составляет 95,45% всей площади под графиком.

Повторите эти операции с 3 сигмами. Должен получиться результат, отражённый на рис. 8:



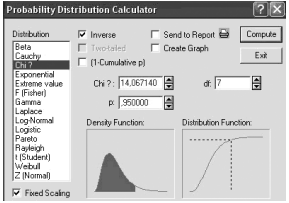
*Рис. 8 – Иллюстрация к правилу трех сигм*

Данные правила действуют при любых значениях среднего и стандартного отклонения нормального распределения.

*Распределение хи-квадрат*

Случайная величина, имеющая распределение хи-квадрат, определяется как сумма квадратов *k* независимых стандартных нормальных величин (величины, имеющие нормальное распределение). Число k в определении хи-квадрата – число степеней свободы. В частном случае, когда k = 1 случайная величина хи-квадрат равна квадрату стандартной нормальной величины. Это распределение имеет только один параметр – число степеней свободы (целое положительное число).

**Шаг 1**. В списке распределений вероятностного калькулятора выберите **Chi ? (хи-квадрат-распределение)** (рис. 9).

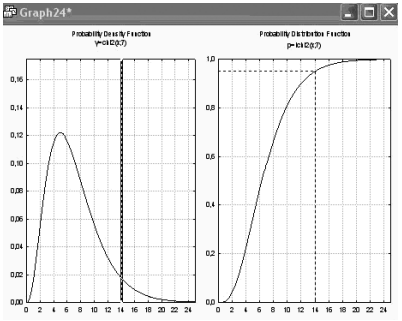


*Рис. 9 – Задание хи-квадрат-распределения в вероятностном калькуляторе*

**Шаг 2.** В строке **df** задайте 7 – число степеней свободы.

**Шаг 3.** В поле **р** задайте 0,95. Нажмите кнопку **Compute (вычислить)**, в строке **Chi ?** вы увидите 14,067140 – квантиль хи-квадрат-распределения с 7 степенями свободы.

**Шаг 4**. Выберите далее опцию **Создать график** и вновь щелкните на кнопке **Compute (вычислить)** и вы увидите график плотности и функции распределения хи-квадрат с 7 степенями свободы (рис. 10).



*Рис. 10 – График плотности и функции распределения случайной величины хи-квадрат с 7 степенями свободы*

Это распределение несимметрично и сосредоточено только на положительной полуоси. Распределение хи-квадрат играет важную роль при исследовании оценки дисперсии нормальной выборки, а также при проверке зависимостей в таблицах сопряженности и в критериях согласия.

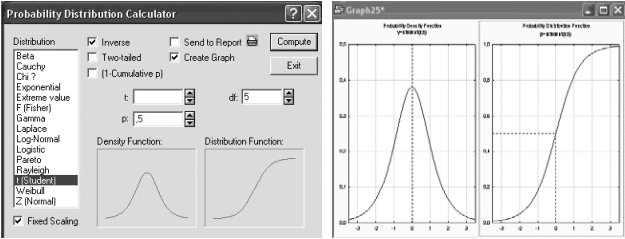
*t-распределение Стьюдента*

t-распределение важно в тех случаях, когда рассматриваются оценки среднего и неизвестна дисперсия выборки. В этом случае используют выборочную дисперсию и t-распределение. t-распределение возникает в таблицах вывода регрессионного анализа. Это одно из важнейших распределений, наряду с нормальным и распределением хи-квадрат. t-распределение с k-степенями свободы сосредоточено на всей действительной оси, симметрично относительно 0. Среднее t-распределения равно 0, дисперсия равна k/(k-2).

**Шаг 1.** В списке распределений вероятностного калькулятора выберите **t (Student) (t-распределение Стьюдента)** (рис. 11).

**Шаг 2.** В строке **df** задайте 5 (число степеней свободы). Пометьте опцию **Create Graph (создать график)**.

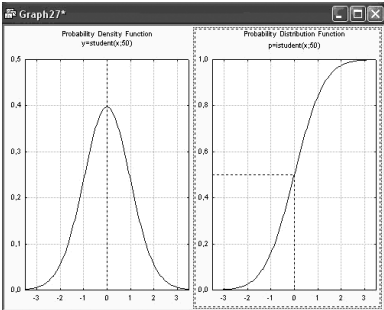
**Шаг 3.** В поле р задайте 0,5. Нажмите кнопку **Compute (вычислить)**, на экране вы увидите следующий график (рис. 12).



|  |  |
| --- | --- |
| *Рис. 11 – Задание t-распределения*  *Стьюдента в вероятностном*  *калькуляторе* | *Рис.12 – Плотность и функция*  *t-распределения Стьюдента*  *с 5 степенями свободы* |

При больших степенях свободы (>30) t-распределение практически совпадает со стандартным нормальным распределением. Плотность t-распределения деформируется при возрастании числа степеней свободы следующим образом: пик увеличивается, «хвосты» более круто идут к 0 (плотность сжимается с боков).

В такой деформации плотности легко убедиться с помощью вероятностного калькулятора. Задайте в поле **df (степень свободы)** значение 50. Нажав кнопку **Compute (вычислить)**, на экране вы увидите график, отображённый на рис. 13.



*Рис. 13 – Плотность и функция распределения Стьюдента с 50 степенями свободы*

Сравнив график плотности распределения Стьюдента с большим числом степеней свободы, например 50, и график плотности стандартного нормального распределения, вы убедитесь, что они очень похожи.

*F-распределение*

F-распределение возникает в регрессионном, дисперсионном и дискриминантном анализе, а также в других видах многомерного анализа данных. Поэтому с ним следует ознакомиться подробнее. Случайная величина, имеющая F-распределение с парой степеней свободы *m*, *n*, определяется как отношение двух независимых случайных величин, имеющих распределение хи-квадрат со степенями свободы *m* и *n* с умножением на нормировочный сомножитель *n/m*. F-распределение сосредоточено на положительной полуоси. Это распределение в отличие от нормального несимметрично.

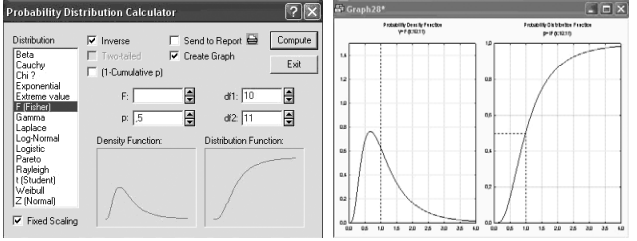
**Задание 3.** Постройте график F-распределения и вычислите его процентные точки.

Решение

**Шаг 1.** В списке распределений вероятностного калькулятора выберите **F (F-распределение)** (рис. 14).

**Шаг 2.** Задайте в поле **df1 (степень свободы 1)** значение 10, в поле **df2 (степень свободы 2)** – значение 11. Пометьте опцию **Create Graph (создать график)**.

**Шаг 3.** В поле **р** задайте 0,5. Нажав кнопку **Compute (вычислить)**, на экране отобразится график (рис. 15).

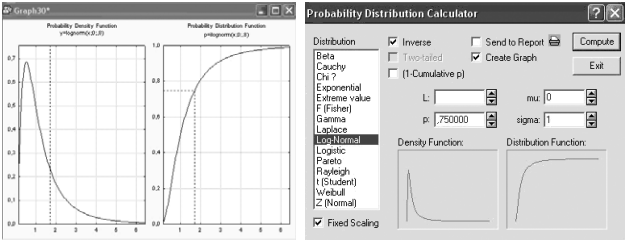


|  |  |
| --- | --- |
| *Рис. 14 – Задание F-распределения в вероятностном калькуляторе* | *Рис.15 – Плотность и функция F-распределения со степенями свободы 10, 11* |

*Логарифмически-нормальное распределение*

Случайная величина X имеет логарифмически-нормальное распределение, если величина Ln (X) является нормальной, т.е. логарифм логарифмически-нормальной величины является нормальной величиной. Так как нормальное распределение описывается двумя параметрами, то и логарифмически-нормальное распределение также имеет два параметра. Плотность распределения имеет одно максимальное значение и несимметрично. График плотности логарифмически-нормального распределения показан на рис.16.

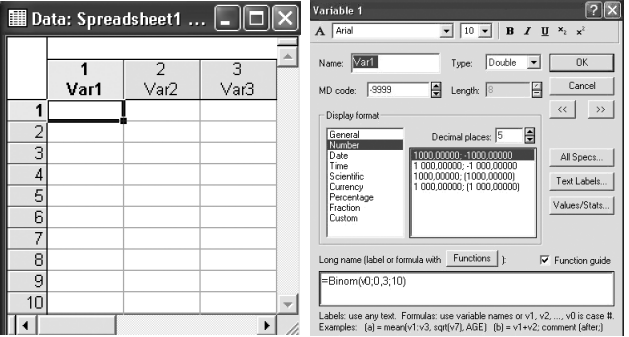
В списке распределений вероятностного калькулятора выберите Log-Normal (логарифмически-нормальное распределение) и задйте параметры, показанные на рисунке (рис. 17).



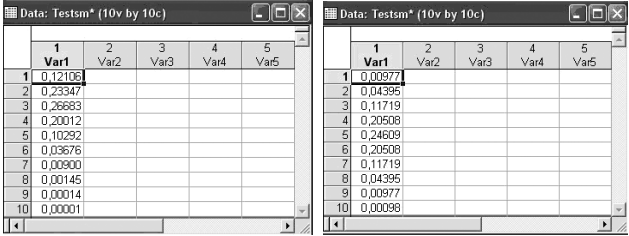
|  |  |
| --- | --- |
| *Рис.16 – Плотность логарифмически-нормального распределения* | *Рис. 17 – Задание логнормального распределения в вероятностном калькуляторе* |

*Биномиальное распределение и игровые задачи*

Создайте пустую электронную таблицу testsm.sta (рис. 18). Дважды щелкните на имени переменной var1 и откройте диалоговое окно спецификации переменной **var1**. В нижней части окна в поле **Long name** запишите формулу, как показано на рис. 19 и нажмите кнопку **OK** в правом верхнем углу окна. Согласно этой формуле программа вычислит вероятности успеха и занесет их в таблицу в значения первой переменной. Теперь таблица примет вид, представленный на рис. 20.



|  |  |
| --- | --- |
| *Рис. 18 – Пустая электронная*  *таблица testsm.sta* | *Рис. 19 – Задание формулы*  *вычисления биномиальных*  *вероятностей* |



|  |  |
| --- | --- |
| *Рис. 20 – Электронная таблица*  *с биномиальными вероятностями (вероятность успеха 0,3, число испытаний 10)* | *Рис. 21 – Электронная таблица*  *с биномиальными вероятностями (вероятность успеха 0,5, число испытаний 10)* |

В данной таблице вероятность успеха – выпадения герба – равна 0,3. Из таблицы видно, что вероятность выпадения ровно одного герба в 10 бросаниях – 0,12106, вероятность выпадения ровно двух гербов в 10 бросаниях – 0,2334 и т. д.

Вероятность успеха можно изменить, сделав ее равной 0,5. Это означает, что бросается симметричная монета и вероятность успеха равна вероятности неудачи (вероятность выпадения герба равна вероятности выпадения решки). Для этого в спецификации переменной **var1** в поле **Long name** измените формулу, вместо 0,3 запишите 0,5 и нажмите кнопку OK. Программа вычислит новые биномиальные вероятности и занесет их в электронную таблицу (рис. 21). Максимальная вероятность в этой таблице приходится на значение 5, что и понятно из соображений симметрии.

*Задача шевалье де Мере*

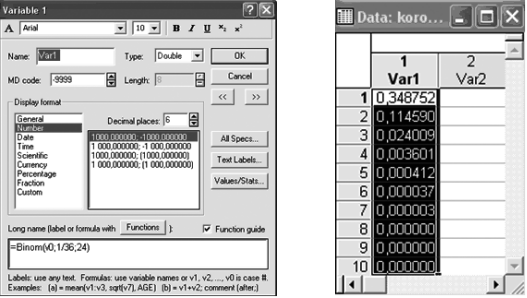
Классическим является пример шевалье де Мере, когда азартный игрок спросил себя: стоит ему ставить на выпадение 2 шестерок одновременно при бросании двух костей 24 раза или нет? Его собственные вычисления показали, что стоит, так как вероятность данного события при 24 бросаниях костей больше 1/2. То это оказалось не так. Знаменитый Блез Паскаль указал ему на заблуждение: оказывается, вероятность данного события 0,49, следовательно, в длинной серии игр, состоящих в 24 подбрасываниях 2 костей, выигрыш происходит лишь в 49%, а не в более чем 50% игр, как рассчитывал де Мере. Шевалье обычно играл всю ночь, и для него было важно, чтобы в более чем половине игр он был в выигрыше.

**Задача.** Решите эту задачу с помощью средств STATISTICA.

Решение

Создайте рабочий файл play.sta. В спецификации переменной **var1** в поле **Long name** запишите формулу, как показано на рис. 22. Число испытаний в задаче шевалье 24, вероятность успеха равна 1/36, потому что с такой вероятностью при бросании 2 костей выпадают шестерки.

Нажмите **OK** и программа вычислит биномиальные вероятности. Результат представлен на рис. 23.



|  |  |  |
| --- | --- | --- |
| *Рис. 22 – Задание формулы*  *биномиального распределения в*  *задаче шевалье* |  | *Рис. 23 – Вероятности выпадения шестерок при 24 бросаниях 2 костей* |

В первом столбце этой таблицы даны последовательно вероятности выпадения 2 шестерок 1 раз, 2 раза, 3 раза и т. д.

По условию задачи спрашивается, стоит ли ставить на выпадение 2 шестерок одновременно при бросании двух костей 24 раза или нет? Нужно вычислить вероятность выпадения по крайней мере одной пары шестерок. Следовательно, все эти вероятности нужно сложить. Сделав это, вы получите ответ к классической задаче. Вероятность выпадения по крайней мере одной пары шестерок при 24 бросаниях пары костей равна 0,49140. Т.е. в длинной серии игр, состоящих из 24 бросаний пары костей, игрок, ставящий на выпадение 2 шестерок одновременно, в среднем устойчиво проигрывает.

**Задание 4.** Будет ли выигрыш, если игра состоит из 25 бросаний, то есть ставка на выпадение пары шестерок в 25 бросаниях.

**Задание 5**. Семь коров пасутся на лугу, равномерно покрытом травой. Коровы не имеют склонности к скоплению, и каждая из них равно предпочитает любую точку луга. Коровы двигаются по лугу совершенно независимо отдна от другой. Площадь луга 10000 м2. Найдите вероятности того, что: ровно 2 коровы окажутся в квадате со стороной 50 м; только 1 корова окажется в данном квадрате; все 7 коров окажутся в данном квадрате.