|  |
| --- |
| Московский Энергетичестий Институт (Технический Университет) |
| Лабораторная работа 5 |
| Критерий хи-квадрат проверки гипотез |
|  |
| **студент Винников Александр** |
| **группа А-14-07** |

|  |
| --- |
| Москва 2010 |

# Работа № 5. Критерий хи-квадрат проверки гипотез

Критерий хи-квадрат Пирсона является весьма общим методом построения тестов для проверки различных гипотез. Рассмотрим исходную схему.

## 1. Описание критерия хи-квадрат

Обозначим:

*A*1*, ..., Am* - *m* возможных исходов некоторого опыта; *p*1*, ..., pm* - вероятности cooтветствующих исходов, ;

*n* - число независимых повторений опыта;

ν1*, ...,* ν*m* - число появлений соответствующих исходов в *n* опытах, ;

*p, ..., p* - гипотетические значения вероятностей, *p>* 0*,* .

Требуется по наблюдениям *ν*1*,...,νm* проверить гипотезу *Н* о том , что вероятности *p*1*, ..., pm* имеют значения *p, ..., p*, т.е.

*Н: pi= p* *, i=*1*, ...,m.*

Оценками для *p*1*, ..., pm* являются = ν1 */n,* *...*, = ν*m/n*. Мерой расхождения между гипотетическими и эмпирическими вероятностями принимается величина

**,

которая с точностью до множителя *n* есть усредненное с весами *p* значение квадрата относительного отклонения значений  от *p**.* Статистика *X2* называется статистикой хи-квадрат Пирсона. Для ее вычисления используются две формулы:

*.* (1)

Поскольку по закону больших чисел*→ pi*при *n* → ∞, то

.

Последняя величина равна 0, если верна *Н*; если же *Н* не верна, то *X2* → ∞.

Процедура проверки гипотезы состоит в том, что если величина *X2* приняла “слишком большое” значение, т.е. если

*X2  ≥ h* , (2)

то гипотеза *Н* отклоняется; если это не так, будем говорить, что наблюдения не противоречат гипотезе. На вопрос, что означает “слишком большое” значение, отвечает

**Теорема К. Пирсона.** Если гипотеза *Н* верна и *pi*0> 0, *i=*1,...,*m*, то при *n→ ∞* распределение статистики *Х2* асимптотически подчиняется распределению хи-квадрат с *m* - 1 степенями свободы, т.е.

*Р{ X2* < *x / H } → Fm-*1*(x) ≡ P{* χ*2m-*1 < *x }.*

Порог *h* выберем из условия: вероятность ошибки первого рода должна быть малой - равной выбираемому значению α - уровню значимости:

*P*{ отклонить *H* / *H* верна} = *P*{ *X 2* ≥ *h / H* } ≅ *P*{χ*2m-*1≥ *h*} = α,

откуда

*h = Q(* 1*-*α*, n -*1*)*  (3)

- квантиль уровня 1-α распределения хи-квадрат с *m* -1 степенями свободы.

Процедура (2) - (3) проверки *Н* может быть записана иначе: гипотеза *Н* отклоняется, если

*P{*χ*2m-*1≥ *X2}* ≤α , (4)

т.е. если мала вероятность получения (при справедливости *Н*) такого же расхождения, как в опыте (т.е. *X2*), или ещё большего. Вероятность слева в (4) называется минимальным уровнем значимости (при любом значении α, большем *P*{*X2m-*1 ≥ *X2*}, гипотеза, очевидно, отклоняется).

**Замечание**. Теорему Пирсона можно применять, если все ожидаемые частоты

*np* ≥ 10, *i=*1*, ...,m;* (5а)

если *m* порядка десяти и более, достаточно выполнения

*np*≥4, *i=*1*, ...,m*. (5б)

Если (5) не выполняется, необходимо некоторые исходы *Аi*объединять

## 3. Проверка гипотезы о типе распределения

Пусть требуется проверить гипотезу о том, что выборка *x*1*, ..., xn* извлечена из совокупности, распределенной по некоторому закону, известному с точностью до *k*-мерного параметра *а=*(*а*1*,...,аk*). Оказываются теоретически обоснованными следующие действия: разобьем весь диапазон наблюдений на *m* интервалов, определим значения *νi* -число наблюдений в *i*-м интервале, получим значение оценки минимизацией (6) или методом максимального правдоподобия, определим вероятности *pi(**)* попадания в *i*-й интервал, вычислим (6) или (8) и примем решение по (7).

**Пример 1. Проверка нормальности.** Проверим гипотезу о нормальном законе распределения размеров головок заклепок, сделанных на одном станке, по выборке объема *n* = 200; измерения приведены в таблице 1 [1, с.15]. Оценками для *а* (среднего) и σ (стандартного отклонения) являются

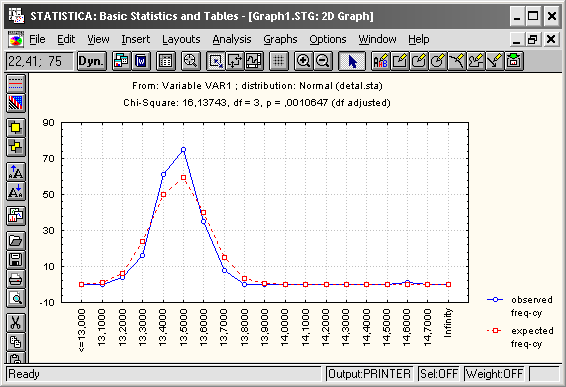
 и .

Таблица 1.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Диаметры 200 головок заклепок, мм | | | | | | | | | |
| 13.39 | 13.33 | 13.56 | 13.38 | 13.43 | 13.37 | 13.53 | 13.40 | 13.25 | 13.37 |
| 13.28 | 13.34 | 13.50 | 13.38 | 13.38 | 13.45 | 13.47 | 13.62 | 13.45 | 13.39 |
| 13.53 | 13.58 | 13.32 | 13.27 | 13.42 | 13.40 | 13.57 | 13.46 | 13.33 | 13.40 |
| 13.57 | 13.36 | 13.43 | 13.38 | 13.26 | 13.52 | 13.35 | 13.29 | 13.48 | 13.43 |
| 13.40 | 13.39 | 13.50 | 13.52 | 13.39 | 13.39 | 13.46 | 13.29 | 13.55 | 13.31 |
| 13.29 | 13.33 | 13.38 | 13.61 | 13.55 | 13.40 | 13.20 | 13.31 | 13.46 | 13.13 |
| 13.43 | 13.51 | 13.50 | 13.38 | 13.44 | 13.62 | 13.42 | 13.54 | 13.31 | 13.58 |
| 13.41 | 13.49 | 13.42 | 13.45 | 13.34 | 13.47 | 13.48 | 13.59 | 13.20 | 14.56 |
| 13.55 | 13.44 | 13.50 | 13.40 | 13.48 | 13.29 | 13.31 | 13.42 | 13.32 | 13.48 |
| 13.43 | 13.26 | 13.58 | 13.38 | 13.48 | 13.45 | 13.29 | 13.32 | 13.24 | 13.38 |
| 13.34 | 13.14 | 13.31 | 13.51 | 13.59 | 13.32 | 13.52 | 13.57 | 13.62 | 13.29 |
| 13.23 | 13.37 | 13.64 | 13.30 | 13.40 | 13.58 | 13.24 | 13.32 | 13.52 | 13.50 |
| 13.43 | 13.58 | 13.63 | 13.48 | 13.34 | 13.37 | 13.18 | 13.50 | 13.45 | 13.60 |
| 13.38 | 13.33 | 13.57 | 13.28 | 13.32 | 13.40 | 13.40 | 13.33 | 13.20 | 13.44 |
| 13.34 | 13.54 | 13.40 | 13.47 | 13.28 | 13.41 | 13.39 | 13.48 | 13.42 | 13.46 |
| 13.28 | 13.46 | 13.37 | 13.53 | 13.43 | 13.30 | 13.45 | 13.40 | 13.45 | 13.40 |
| 13.33 | 13.39 | 13.56 | 13.46 | 13.26 | 13.35 | 13.42 | 13.36 | 13.44 | 13.41 |
| 13.43 | 13.51 | 13.51 | 13.24 | 13.34 | 13.28 | 13.37 | 13.54 | 13.43 | 13.35 |
| 13.52 | 13.23 | 13.48 | 13.48 | 13.54 | 13.41 | 13.51 | 13.44 | 13.36 | 13.36 |
| 13.53 | 13.44 | 13.69 | 13.66 | 13.32 | 13.26 | 13.51 | 13.38 | 13.46 | 13.34 |

***Выполнение***

Результаты измерения диаметров заклепок занесем в таблицу с одним столбцом (*d*) и 200 строками. Выберем распределение из семейства нормальных с параметрами среднее - 13.42, дисперсия - 0*.*01*8,* группирование с числом групп 19. Сравним графически наблюдаемые и ожидаемые частоты:



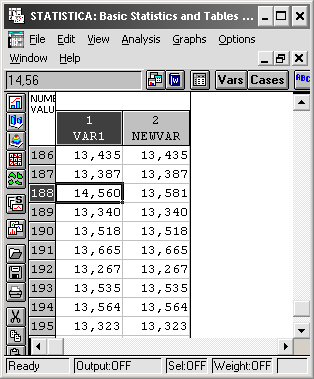
В таблице приведено значение статистики (8) *Chi-Square:* 16.14, количество степеней свободы *df* = 3, которое получилось при объединении интервалов для выполнения условий (5): *f* = 6 - 1 - 2 = 3. Приведено значение вероятности

Р⎨ χ2 3 ≥ 16.14⎬ = *р* = 0.001.

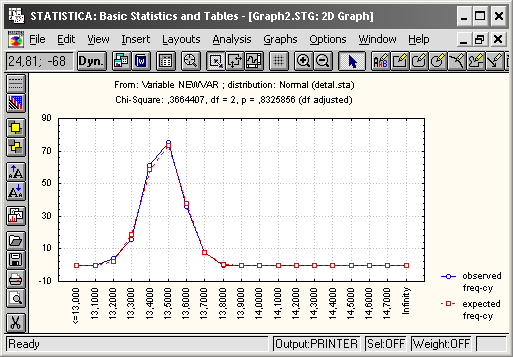
последнее означает, что если гипотеза верна, вероятность получить 12.00 слишком мала, чтобы поверить в нормальность. Гипотезу о нормальности отклоняем.

Если посмотреть гистограмму наблюдений, видно, что в выборке имеется одно аномальное значение 14.56 (№ 188), которое могло появиться в результате какой-либо ошибки. Удалим его и снова проверим гипотезу. Удаление одного наблюдения, если оно типично, не может изменить характеристики совокупности из 200 элементов; если же изменение происходит, следовательно, это наблюдение типичным не является и должно быть удалено.

Чтобы не портить исходные данные, продублируем их в новый столбец, например, *dc*, и удалим аномальное наблюдение:



Повторим проверку гипотезы для “цензурированной” выборки и убедимся в том, что наблюдения не противоречат гипотезе о нормальности:



## 4. Примеры проверки простой гипотезы о распределении

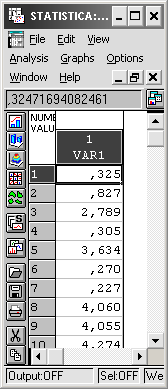
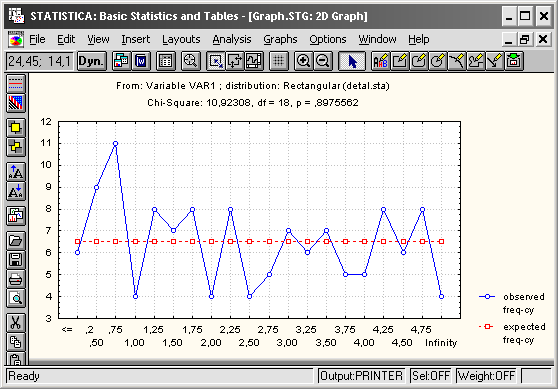
**Пример 2.** Проверим генератор случайных чисел. Сгенерируем выборку заданного объема с заданным законом распределения, и по полученным результатам проверим гипотезу о согласии данных с этим распределением.

Вариант 1

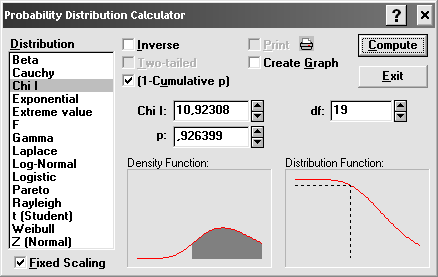
|  |  |
| --- | --- |
| Распределение  Объем | *R*[0, 5]  130 |

***Выполнение***

Аналогично предыдущему примеру, сгенерируем значения и выберем из семейства распределений подходящее для оценивания(групп 20, среднее 2,5):



Приводимый результат для уровня значимости *р* не соответствует рассматриваемому случаю, так как число степеней свободы *d.f.* должно быть равным  *m* -1; пакет же указывает с учетом числа оцениваемых параметров. Получим нужное значение для *р*:



**Пример 3.** В опытах по генетике Мендель наблюдал частоты появления различных видов семян, получаемых при скрещивании гороха с круглыми желтыми и с морщинистыми зелеными семенами [2]. Частоты приведены в таблице 3 вместе с теоретическими вероятностями.

Таблица 3. Частоты видов семян.

|  |  |  |
| --- | --- | --- |
| Семена | Наблюдаемая  частота, ν*i* | Теоретическая  вероятность, *pi* |
| Круглые и желтые  Морщинистые и желтые  Круглые и зеленые  Морщинистые и зеленые | 315  101  108  32 | 9/16  3/16  3/16  1/16 |
| Сумма | *n* = 556 |  |

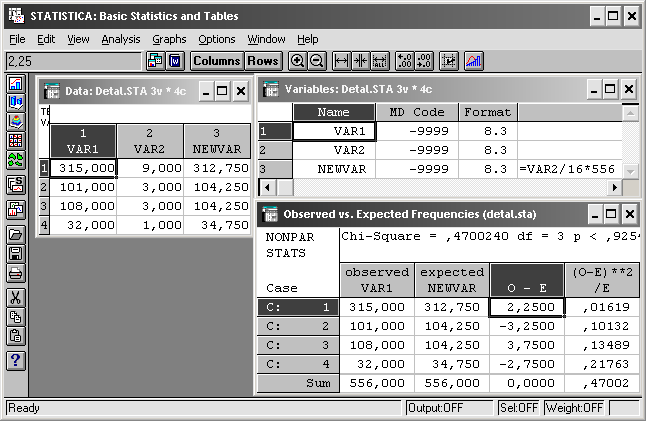
Формула (1) дает *X*2 = 0.47. При числе степеней свободы *m*-1 = 3

*P*{**≥ 0.47 } = 0.92,

так что между наблюдениями и теорией имеется очень хорошее согласие: критерий с любым уровнем значимости α ≤ 0.92 не отвергал бы эту гипотезу .

***Выполнение***

Перенесём данные и воспользуемся процедурой *Observed versus expected* (наблюдаемые частоты против ожидаемых):



Над таблицей наблюдаем значение Хи-квадрат, число степеней свободы и вероятность правильности гипотезы H – 0,925425

## 5. Проверка гипотезы о независимости признаков (таблица сопряженности признаков)

**Пример 4.** Данные [2], собранные по ряду школ, относительно физических недостатков школьников (*P*1, *P*2, *P*3 - признак *А*) и дефектов речи (*S*1, *S*2, *S*3 - признак *В*) приведены в таблице 4. В таблице 5 даны частоты.

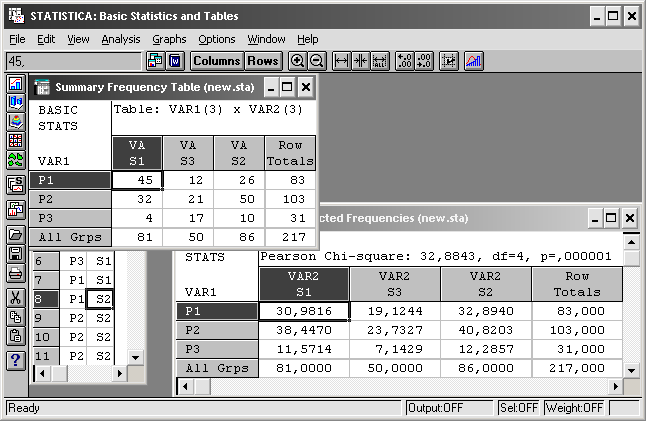
Для проверки гипотезы о независимости этих двух признаков вычислим статистику (11):  = 34.88; число степеней свободы *f* = (3-1)×(3-1) = 4; минимальный уровень значимости

;

это значит, что при независимых признаках вероятность получить значение такое же, как в опыте или большее, меньше 0.001, и потому гипотезу о независимости следует отклонить.

***Выполнение***

Образуем таблицу с двумя столбцами (*P* и  *S*) и 217 строками. Получим результаты таблиц сопряженности(слева) и ожидаемые или теоретические частоты(справа):



, и потому гипотезу о независимости следует отклонить.

## 6. Проверка гипотезы об однородности выборок

Пусть имеется *m* выборок объемами *n*1,..., *nm*, извлеченных из различных совокупностей. Измеряемая величина в каждой из выборок может иметь *k* уровней *B*1, ..., *B*k. Требуется проверить гипотезу о том, что исходные совокупности распределены одинаково. Обозначим *νij* - число наблюдений в *i*-й выборке, имеющих уровень *B*j,  . Имеем таблицу *m×k* наблюдений налогично предыдущему пункту 5. Можно показать, что для проверки гипотезы справедлива процедура (11) - (12).

**Пример 5.** Имеются данные [3] о наличии примесей серы в углеродистой стали, выплавляемой двумя заводами (см. таблицу 6).

Таблица 6. Число плавок

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Содержание серы, 10-2 % | | | | |
|  | 0÷2 | 2÷4 | 4÷6 | 6÷8 | Сумма |
| Завод 1  Завод 2 | 82  63 | 535  429 | 1173  995 | 1714  1307 | 3504  2794 |
| Сумма | 145 | 964 | 2168 | 3021 |  |

Проверим гипотезу о том, что распределения содержания серы (нежелательный фактор) одинаковы на этих заводах.

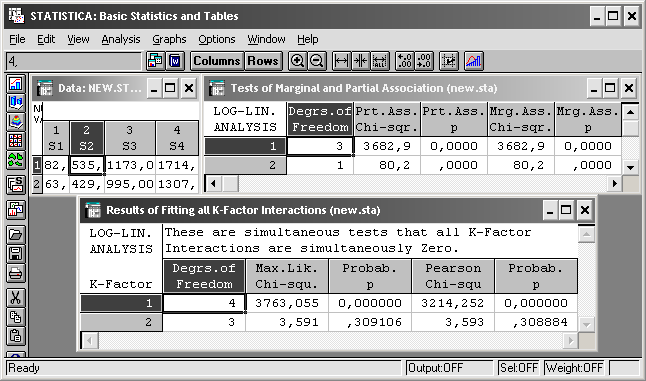
По (11) находим:  = 3.39. Число степеней свободы f = (2-1)×(4-1) = 3; квантиль уровня 0.95

*h = Q(*0.95, 3*) =* 7.8.

Полученное нами из опыта значение 3.39 лежит в области допустимых значений, и потому у нас нет оснований считать, что содержание серы в стали этих заводов имеют различные распределения.

***Выполнение***

Образуем таблицу 2 × 4, в которую занесем данные; столбцы назовем, например, S1 ÷ S4 (сера) , а строки - Z1, Z2 (заводы).



В таблице *Results of Fitting..*. в последней строке столбца получаемзначение хи-квадрат = 3.59, число степеней свободы *f* = 3, и уровень значимости *p* = 0.31. поскольку эта вероятность не мала (не является значимой), гипотезу об одинаковом распределении содержания серы в металле на двух заводах можно принять (наблюдения этому не противоречат).

## 7. Персональное задание

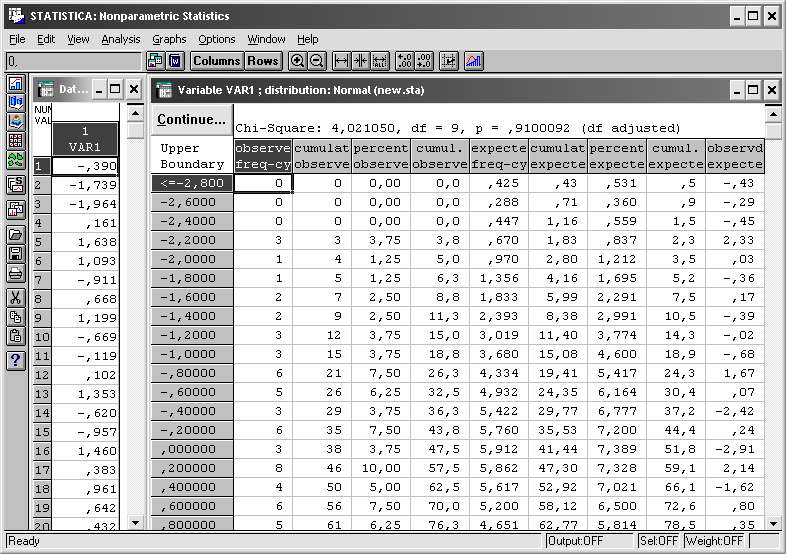
1. Проверить гипотезу о типе распределения на основе сгенерированной по заданному в таблице 7 закону выборке объема *n*. Проверить три гипотезы: о нормальности, о равномерности и о показательности**.**

Таблица 7. Исходные данные

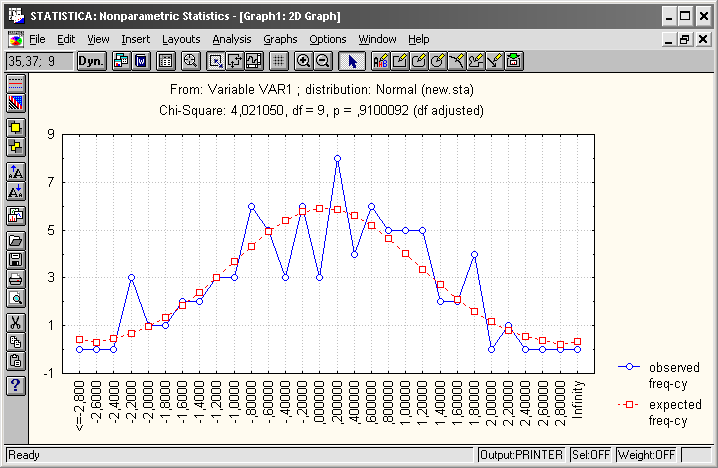
|  |  |
| --- | --- |
| Вариант | 1 |
| Распределение  Объем | *N*[0, 1]  80 |

***Выполнение***

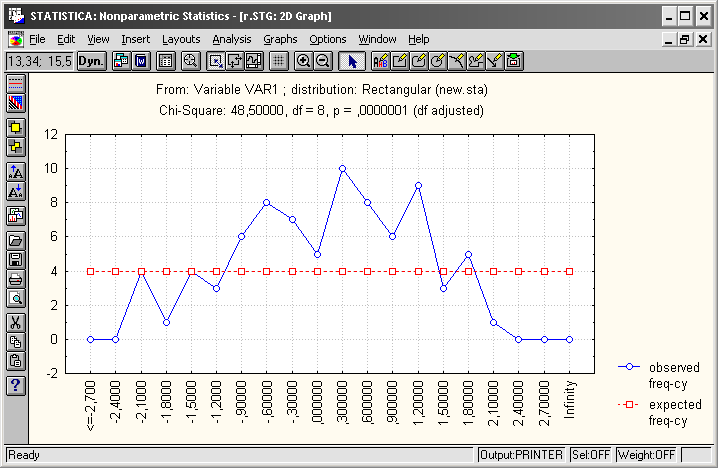
Сгенерируем выборку по нормальному закону, сгруппируем значения в 30 групп:

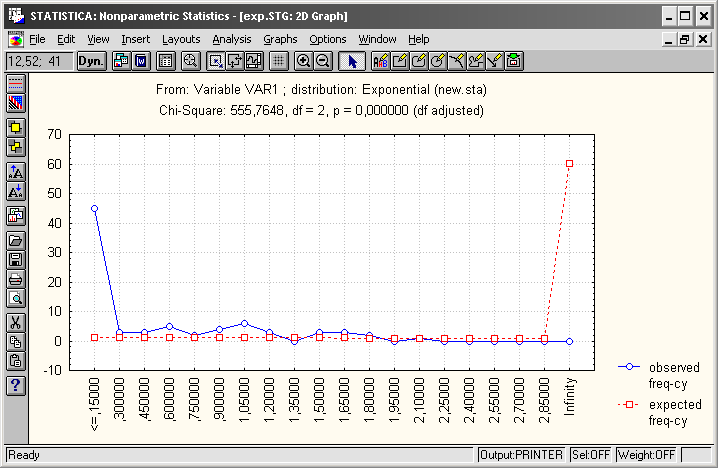


Получаем уровень значимости 0.91. Гипотезу о нормальности принимаем.



Гипотезу о равномерности или показательности выборки отклоняем:





По показательному закону выборка не может быть распределена вовсе, поскольку имеются отрицательные значения.

2. Проверить гипотезу об однородности трех выборок.

Сгенерировать три выборки объемами *n*1 = 180, *n*2 = 100, *n*3 = 120для заданного в таблице 8 распределения. Провести их группирование на 8 10 интервалах. Сделать все для 2-х вариантов:

а) параметры одинаковы;

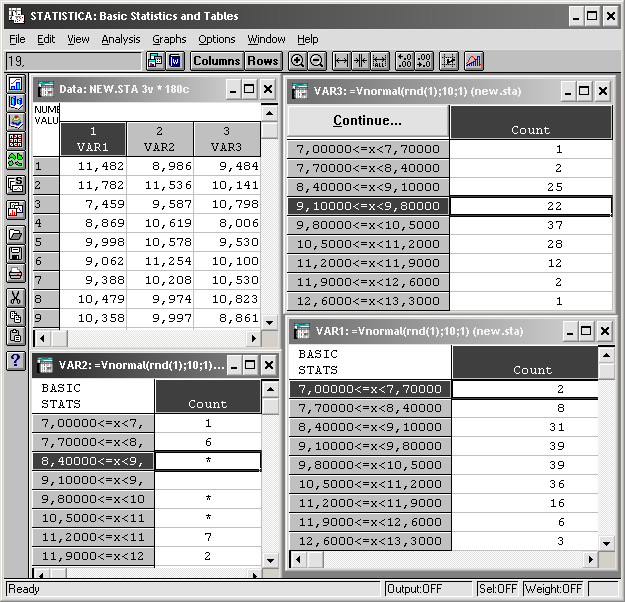
б) параметры различны.

Таблица 8. Исходные данные.

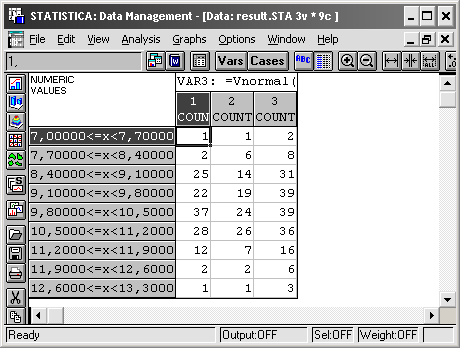
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| N | Тип | а) | б) | | |
|  |  | *a*1 *= a2 = a3* | *a*1 | *a2* | *a3* |
| 1 | *N(a,* 1*)* | 10  *n*1 = 180, *n*2 = 100, *n*3 = 130 | 9.8  180 | 10  100 | 11.2  120 |

***выполнение***

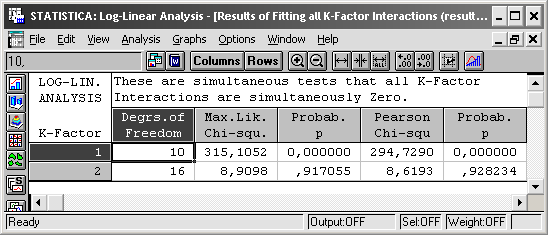
а) Сгенерируем выборки по нормальному закону с параметрами 10; 1 объёма 180, 100, 130 и сгруппируем их в 9 групп:



Объединим группы в одну таблицу

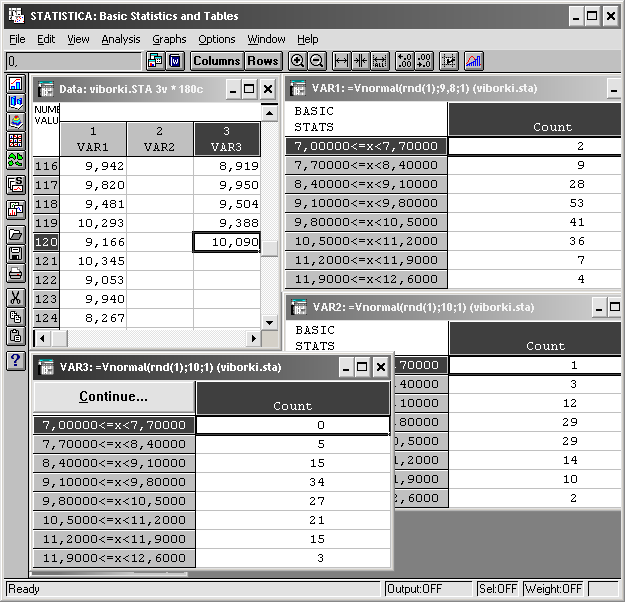


В результирующей таблице получаем:

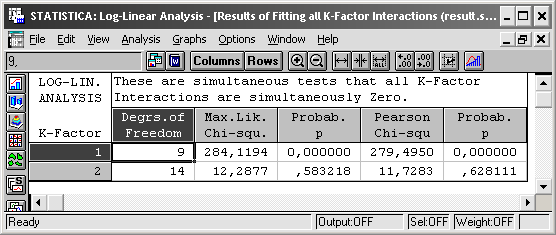


Значение хи-квадрат = 8,909 число степеней свободы *f* = 16, и уровень значимости *p* = 0.917. поскольку эта вероятность не мала, гипотезу об одинаковом распределении можно принять (вернее, наблюдения этому не противоречат).

б) Сгенерируем выборки с разными параметрами и сгруппируем значения:



В результате получаем:



Количество степенией свободы = 14, Значение статистики = 12,2877; Уровень значимости =0,5832. Однородность меньше, чем в пункте а).