Министерство образования и науки Российской Федерации

Новосибирский государственный технический университет

Кафедра ТПИ

**Курсовой проект**

по дисциплине «Компьютерные технологии анализа данных и исследования статистических закономерностей»

Факультет: ПМИ

Группа: ПММ-71

Вариант: 15

Студент: Кочнев А. В.

Преподаватель: Лемешко Б. Ю.

Семенова М.А.

Новосибирск

2017

**Оглавление**

[**Введение** 3](#_Toc501778360)

[**Постановка задачи** 4](#_Toc501778361)

[**Результаты численных экспериментов** 7](#_Toc501778362)

[**Пример анализа реальных данных** 12](#_Toc501778363)

[**Список используемых источников** 15](#_Toc501778364)

# **Введение**

Равномерно распределенные случайные величины в реальных данных встречаются достаточно редко, но они играют особую роль в генерации случайных величин, распределенных по другим законам. Следовательно, от качества генератора равномерно распределенных случайных величин зависят генераторы случайных чисел, распределенных по другим законам. Существует множество критериев, с помощью которых можно проверить гипотезу о принадлежности выборки равномерному закону.

Все критерии применяемые для проверки равномерности можно разделить на несколько групп [1]:

* критерии, чьи статистики основаны на разнице элементов вариационного ряда
* критерии, использующие разности оценок порядковых статистик исследуемой выборки, и математические ожидания этих статистик
* критерии основанные на оценках энтропии
* непараметрические критерии согласия, применяемые для проверки равномерности
* критерий -Пирсона

Критерий Шермана, с различными модификациями статистики относится к критериям первой группы. В данной работе рассматривается критерий с нормализованной и модифицированной статистикой.

Целью выполнения данного курсового проекта является:

Знакомство с современными тенденциями развития аппарата прикладной математической статистики и состоянием программного обеспечения задач статистического анализа. Освоение методов статистического моделирования как средства исследования и развития аппарата прикладной математической статистики. Исследование особенностей методов проверки статистических гипотез. Закрепление навыков проведения самостоятельных исследований.

# **Постановка задачи**

Равномерное распределение случайной величины X на интервале [0,1] будем обозначать как . Функция распределения вероятностей равномерного закона имеет вид , при x∈[0,1]. Для случайной величины, равномерно распределённой на интервале [a,b], функция распределения имеет вид при x∈[a,b].

При проверке гипотезы о принадлежности наблюдаемой случайной величины равномерному закону простая проверяемая гипотеза имеет вид H0 : X ∈ или H0 : X ∈ , где a и b известны. Эту же гипотезу можно записать как H0 : , x∈[0,1] или H0 :, x∈[a,b].

Проверяемая гипотеза будет сложной, если по данной выборке находится и область определения равномерной случайной величины.

Пусть X1,X2,…,Xn – выборка независимых наблюдений случайной величины X .

Статистики критериев первой группы предусматривают использование разностей последовательных значений вариационного ряда

, где ,, n – объем выборки.

К ним и относится рассматриваемый далее критерий Шермана [2].

Статистика данного критерия имеет вид:

,

где элементы вариационного ряда, построенного по выборке  объемом n.

Критерий правосторонний: гипотеза H0 отклоняется при больших значениях статистики .

Суть критерия заключается в том, что разница между соседними элементами вариационного ряда должна быть близка к длине отрезка обратно пропорциональной объему выборки, что обеспечивает равномерность распределения элементов вариационного ряда.

Критические значения статистики критерия предоставлены в работах [2, 3, 4].

При справедливости H0 и больших объемах выборок n, статистика критерия приближенно подчиняется нормальному распределению.

При проверке гипотез о равномерности также может использоваться нормализованная статистика:

где D и E являются дисперсией и математическим ожиданием. При нормализованная статистика стремится к стандартному нормальному закону [1]. Отличием статистики от нормального закона можно пренебрегать при .

Также существует модификация статистики, задаваемая выражением:

где

Распределение модифицированной статистики еще быстрее сходится к стандартному нормальному закону.

Если объем выборки позволяет использовать нормальную аппроксимацию, то оценка достигнутого уровня значимости (p-value), задаваемая соотношением:

определяется как:

где - значение используемой статистики, вычесленное по анализируемой выборке. Проверяемая гипотеза не отклоняется если для заданной вероятности ошибки первого рода.

**Аналитический обзор**

Рассматриваемый критерий был предложен Шерманом, в 1950 году в его статье «A random variable related to the spacing of sample values», там же была предложена нормализованная статистика критерия [1]. В 1957 году была опубликована статья «Percentiles of wn statistic», в которой были приведены критические значения статистик критерия [4].

В работе 1954 года, Бартоламью предложил модификацию критерия Шермана, сходящуюся к стандартному нормальному распределению при меньших объемах выборки [5].

Достаточно подробно мощность критерия и предельные распределения статистик были рассмотрены в [1]

# **Результаты численных экспериментов**

Сгенерируем выборку, равномерно распределенную на отрезке [0, 1], после чего рассчитаем значения статистик (нормализованной и модифицированной), определим достигнутые уровни значимости и сравним с заданным уровнем значимости, на основании чего сделаем вывод о справедливости гипотезы. Достигнутые уровни значимости будем рассчитывать как , а также с помощью метода Монте-Карло, где количество повторений составляет 16600.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | | | |
| N |  |  | p-value | p-value (Монте-Карло) | Результат |
| 20 | 0.3329 | -0.4933 | 0.6891 | 0.554 | Не отвергается |
| 50 | 0.3851 | 0.6149 | 0.2693 | 0.205 | Не отвергается |
| 100 | 0.3482 | -0.7370 | 0.7694 | 0.722 | Не отвергается |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | | | |
| N |  |  | p-value | p-value (Монте-Карло) | Результат |
| 20 | 0.3329 | -0. 4891 | 0. 6876 | 0. 5555 | Не отвергается |
| 50 | 0.3851 | 0. 6155 | 0. 2690 | 0. 2057 | Не отвергается |
| 100 | 0.3482 | -0. 7373 | 0. 7695 | 0. 7140 | Не отвергается |

Сгенерируем выборку из стандартного нормального закона, проверим гипотезу о равномерности сгенерированной выборки.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | | | |
| N |  |  | p-value | p-value (Монте-Карло) | Результат |
| 20 | 2.2006 | 34.910 | 0 | 0 | Отвергается |
| 50 | 3.2547 | 85.1054 | 0 | 0 | Отвергается |
| 100 | 3. 2804 | 120.6240 | 0 | 0 | Отвергается |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | | | |
| N |  |  | p-value | p-value (Монте-Карло | Результат |
| 20 | 2.2006 | 34.9415 | 0 | 0 | Отвергается |
| 50 | 3.2547 | 85.1075 | 0 | 0 | Отвергается |
| 100 | 3. 2804 | 120.612 | 0 | 0 | Отвергается |

Проверка рассмотренных гипотез в ISW:

|  |  |  |  |
| --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | |
| N |  | p-value | Результат |
| 20 | 0.351765 | 0.559 | Не отвергается |
| 50 | 0.386595 | 0.262 | Не отвергается |
| 100 | 0.352526 | 0.702 | Не отвергается |

|  |  |  |  |
| --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | |
| N |  | p-value | Результат |
| 20 | 2.69876 | 0 | Отвергается |
| 50 | 3.80768 | 0 | Отвергается |
| 100 | 4.06648 | 0 | Отвергается |

Также проверим гипотезы для выборок малого объема, воспользовавшись критическими точками, предоставленными в [1, 3].

Для равномерного распределения на отрезке [0, 1]:

|  |  |  |  |
| --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | |
| N |  | Критическая точка, | Результат |
| 5 | 0.4403012424 | 0.462, 0.1 | Не отвергается |
| 10 | 0.235129973493 | 0.445, 0.1 | Не отвергается |
| 15 | 0.368494119934 | 0.434, 0.1 | Не отвергается |

Для стандартного нормального распределения:

|  |  |  |  |
| --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | |
| N |  | Критическая точка, | Результат |
| 5 | 0.462506471231 | 0.502, 0.05 | Не отвергается |
| 10 | 2.42250091197 | 0.527, 0.01 | Отвергается |
| 15 | 2.94688094994 | 0.502, 0.01 | Отвергается |

Проверим гипотезу о согласии с нормальным распределением, “вписав” его в отрезок [0, 1], со следующими параметрами N(0.5, 1.6). Определим начиная с какого объема выборки критерий отвергнет гипотезу с уровнем значимости 0.01.

|  |  |  |  |
| --- | --- | --- | --- |
| *Критерий Шермана, статистика* | | | |
| N |  | p-value | Результат |
| 25 | 1.3503057920612769 | 0.088458957611654676 | Не отвергается |
| 50 | 2.2639383507099251 | 0.011788953386230872 | Не отвергается |
| 75 | 2.2916025104462965 | 0.01096429524531439 | Не отвергается |
| 100 | 4.2513514695125245 | 1.06242203657958e-05 | Отвергается |
| 125 | 6.6697336211896463 | 1.2813416994106319e-11 | Отвергается |

Построим графики распределений статистик критериев при верной гипотезе H0, для различных модификаций статистик и объемов выборок. Проверим сложную гипотезу о согласии с нормальным распределением, для различных объемов выборок и модификаций статистик критерия.

C:\Users\Андрей\Downloads\normalized.emf

*Рис 1. Зависимость распределения нормализованной статистики критерия от объема выборки*

C:\Users\Андрей\Downloads\modified.emf

*Рис 2. Зависимость распределения модифицированной статистики критерия от объема выборки*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
| N=50 | | | |
| Критерий согласия | S-критерия | p-value | Оценки параметров |
| Колмогорова | 1.2451 | 0.00120 | N(-0.3180,1.0036) |
| Омега-малое | 0.3353 | 0.00012 |
| Омега-большое | 2.0063 | 2.67175e-05 |
| N=50 | | | |
| Колмогорова | 1.13052 | 0.00465 | N(-0.1949,0.9884) |
| Омега-малое | 0.32886 | 0.000153 |
| Омега-большое | 2.14962 | 8.111261 |
| N=200 | | | |
| Колмогорова | 0.814292 | 0.1196 | N(-0.1052,0.9956) |
| Омега-малое | 0.108530 | 0.08590 |
| Омега-большое | 0.64738 | 0.09141 |
| N=2000 | | | |
| Колмогорова | 0.71832 | 0.2601 | N(-0.0221,0.9972) |
| Омега-малое | 0.063369 | 0.3443 |
| Омега-большое | 0.342536 | 0.49384 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
| N=50 | | | |
| Критерий согласия | S-критерия | p-value | Оценки параметров |
| Колмогорова | 0.5500 | 0.6954 | N(-0.3160,1.0161) |
| Омега-малое | 0.0441 | 0.6048 |
| Омега-большое | 0.2675 | 0.6999 |
| N=50 | | | |
| Колмогорова | 0.78995 | 0.14742 | N(-0.1945,0.9923) |
| Омега-малое | 0.07579 | 0.23456 |
| Омега-большое | 0.49583 | 0.21551 |
| N=200 | | | |
| Колмогорова | 1.01928 | 0.01601 | N(-0.1062,0.9969) |
| Омега-малое | 0.17355 | 0.01309 |
| Омега-большое | 0.8977 | 0.02245 |
| N=2000 | | | |
| Колмогорова | 0.68092 | 0.33918 | N(-0.0259,0.9972) |
| Омега-малое | 0.06588 | 0.31873 |
| Омега-большое | 0.36913 | 0.43084 |

# **Пример анализа реальных данных**

В качестве реальных данных возьмем идентификаторы последних открывшихся заведений относящихся к категории «Кафе» из баз данных «2 Гис». Так-как критерий применим только для определения того, равномерна ли выборка на отрезке [0, 1], выполним с данными идентификаторами следующие преобразования:

- вычтем из каждого идентификатора значение наименьшего из них, сдвинув шкалу в 0

- нормируем шкалу, разделив значения идентификаторов на больший из них

В результате получим выборку из 37 элементов со следующими значениями (данные действительны на 23.12.17):

0.8604378 0. 0.34690491 0.29615863 0.10976389 0.11115405  
0.76859714 0.0359202 0.65095386 0.79578036 0.26561745 0.54631635  
0.35308006 0.19011048 0.70165372 0.58349698 0.07323465 0.09717325  
0.11101203 0.52175231 0.07309809 0.30881891 0.58487211 0.82662743  
0.26072594 0.89883789 0.86593836 0.94919909 0.94735146 0.3650876  
0.56608994 0.68226522 0.90065275 0.92932172 0.94942304 0.95066026  
1.

В результате проверки гипотезы критерием Шермана получим следующие результаты, также проверив гипотезу о равномерности критерием Хегази-Грина (так как его мощность достаточно высока)[6]:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Значение статистики | P-value | P-value (Монте-Карло) |
|  | 0.4685 | 0.3197 | 0.2398 |
|  | 0.4540 | 0.3248 | 0.2373 |
| Хегази-Грина | 0.0572137 | 0.328 | - |

**Выводы**

В результате проведения численных экспериментов удалось убедится в следующем:

* Статистика критерия при больших объемах выборки действительно стремится к N(0, 1)
* Модифицированная статистика критерия сходится к нормальному закону при меньших объемах выборки
* Распределение статистики критерия смещенно влево, при том что критерий правосторонний, что сказывается на мощности критерия

# **Список используемых источников**

1. Лемешко Б.Ю, П.Ю Блинов КРИТЕРИИ ПРОВЕРКИ ОТКЛОНЕНИЯ РАСПРЕДЕЛЕНИЯ ОТ РАВНОМЕРНОГО ЗАКОНА. - Новосибирск: НГТУ, 2015.
2. Sherman B. A random variable related to the spacing of sample values / B. Sherman // The Annals of Mathematical Statistics. – 1950. – V.21, №3. – P. 339-361.
3. Кобзарь А. И. Прикладная математическая статистика. Для инжене- ров и научных работников / А. И. Кобзарь. – M : ФИЗМАТЛИТ, 2006. – 816 с.
4. Sherman B. Percentiles of the wn statistic / B. Sherman // The Annals of Mathematical Statistics. – 1957. – V.28, №1. – P. 257-261.
5. Bartholomew D. J. Note on the use of Sherman`s statistics as a test fir randomness // Biometrika. 1954. V. 37. P. 308-312
6. Лемешко Б.Ю, Блинов П.Ю. ОБЗОР СВОЙСТВ КРИТЕРИЕВ РАВНОМЕРНОСТИ // АПЭП, XII МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ. 2014.

**ПРИЛОЖЕНИЕ. ИСХОДНЫЙ КОД ОСНОВНЫХ МОДУЛЕЙ**

*Модуль критерия Шермана*

**import** numpy **as** np  
**from** scipy.stats **import** norm  
**from** params\_by\_sample\_size **import** get\_d, get\_m  
**import** numpy.random **as** n\_rand  
**from** critical\_points **import** sherman\_critical\_points, get\_column\_by\_alpha  
  
  
**def** sherman\_w(sample):  
 variation\_series = np.sort(sample)  
 n = len(variation\_series)  
  
 one\_by\_n = 1/(n+1)  
 w = 0  
 i = 0  
  
 **while** i < len(sample):  
 u\_cur = variation\_series[i]  
 u\_prev = 0  
 **if** i != 0:  
 u\_prev = variation\_series[i-1]  
  
 w += 0.5 \* np.abs(u\_cur - u\_prev - one\_by\_n)  
 i = i + 1  
  
 **return** w  
  
  
*# Normalized Sherman criteria***def** sherman\_w\_normalized(sample):  
 n = len(sample)  
 m = get\_m(n)  
 d = get\_d(n)  
  
 **return** (sherman\_w(sample) - m) / np.sqrt(d)  
  
  
**def** sherman\_w\_normalized\_p\_val(w, m\_carlo\_count, sample\_size):  
 i = 0  
 count = 0  
 **while** i < m\_carlo\_count:  
 y = n\_rand.sample((sample\_size),)  
 w\_cur = sherman\_w\_normalized(y)  
 **if** w\_cur > w:  
 count = count + 1  
 i = i + 1  
  
 **return** count/m\_carlo\_count  
  
  
**def** sherman\_check\_normalized(sample, alpha, m\_carlo\_count = 16600):  
 w = sherman\_w\_normalized(sample)  
 monte\_carlo\_p\_val = sherman\_w\_normalized\_p\_val(w, m\_carlo\_count, len(sample))  
  
 **if** len(sample) >= 20:  
 p\_value = 1.0 - norm.cdf(w)  
 **return** [w, p\_value, p\_value > alpha, monte\_carlo\_p\_val, monte\_carlo\_p\_val > alpha]  
  
 **else**:  
 row = len(sample) - 4  
 column = get\_column\_by\_alpha(alpha)  
 critical\_point = sherman\_critical\_points[row, column]  
 **return** [w, **None**, w < critical\_point, monte\_carlo\_p\_val, monte\_carlo\_p\_val > alpha]  
  
  
*# Modified Sherman criteria***def** sherman\_w\_modified(sample):  
 n = len(sample)  
 w = sherman\_w(sample)  
 bottom = (0.2431/np.sqrt(n)) \* (1.0 - 0.605/n)  
 top = w - 0.3679 \* (1.0 - (1.0/(2.0 \* n)))  
 v = top/bottom  
  
 **return** v - ((0.0995/np.sqrt(n)) \* ((v \* v ) - 1.0))  
 *# return v***def** sherman\_w\_modified\_p\_val(w, m\_carlo\_count, sample\_size):  
 i = 0  
 count = 0  
 **while** i < m\_carlo\_count:  
 y = n\_rand.sample((sample\_size),)  
 w\_cur = sherman\_w\_modified(y)  
 **if** w\_cur > w:  
 count = count + 1  
 i = i + 1  
  
 **return** count/m\_carlo\_count  
  
  
**def** sherman\_check\_modified(sample, alpha, m\_carlo\_count = 16600):  
 w = sherman\_w\_modified(sample)  
 monte\_carlo\_p\_val = sherman\_w\_modified\_p\_val(w, m\_carlo\_count, len(sample))  
  
 **if** len(sample) >= 20:  
 p\_value = 1.0 - norm.cdf(w)  
 **return** [w, p\_value, p\_value > alpha , monte\_carlo\_p\_val, monte\_carlo\_p\_val > alpha]  
  
 **else**:  
 row = len(sample) - 4  
 column = get\_column\_by\_alpha(alpha)  
 critical\_point = sherman\_critical\_points[row, column]  
 **return** [w, **None**, w < critical\_point , monte\_carlo\_p\_val, monte\_carlo\_p\_val > alpha]

*Распределение статистик критерия*

**def** estimate\_params\_normalized():  
 samples\_sizes = [20, 50, 100, 200, 1000, 2000]  
 samples\_count = 10000  
  
 sample\_size\_num = 0  
  
 **while** sample\_size\_num < len(samples\_sizes):  
 sample\_size = samples\_sizes[sample\_size\_num]  
 file = open(**'./output/%d\_%d\_%s.dat'** % (sample\_size, samples\_count, **'normalized'**), **'w'**)  
 file.write(**'Normalaized %d\n0 %d\n'** % (sample\_size, samples\_count))  
 sample\_index = 0  
  
 **while** sample\_index < samples\_count:  
 n\_rand.seed(sample\_index)  
 sample = n\_rand.sample(sample\_size)  
 file.write(str(sherman\_w\_normalized(sample)))  
 file.write(**'\n'**)  
 sample\_index = sample\_index + 1  
 file.close()  
  
 sample\_size\_num = sample\_size\_num + 1  
  
  
**def** estimate\_params\_modified():  
 samples\_sizes = [20, 50, 100, 200, 1000, 2000]  
 samples\_count = 10000  
  
 sample\_size\_num = 0  
  
 **while** sample\_size\_num < len(samples\_sizes):  
 sample\_size = samples\_sizes[sample\_size\_num]  
 file = open(**'./output/%d\_%d\_%s.dat'** % (sample\_size, samples\_count, **'modified'**), **'w'**)  
 file.write(**'Modified %d\n0 %d\n'** % (sample\_size, samples\_count))  
 sample\_index = 0  
  
 **while** sample\_index < samples\_count:  
 n\_rand.seed(sample\_index)  
 sample = n\_rand.sample(sample\_size)  
 file.write(str(sherman\_w\_modified(sample)))  
 file.write(**'\n'**)  
 sample\_index = sample\_index + 1  
 file.close()  
  
 sample\_size\_num = sample\_size\_num + 1