Отчёт по лабораторной №1

«Применение критерия согласия Пирсона»

по дисциплине

«Стохастические модели и анализ данных»

Выполнила: Кацман Н.И.

группа: 3640102/90201

Проверил: к.ф.-м.н., доцент

Баженов А.Н.

Оглавление

[Постановка задачи 3](#_Toc66742380)

[Теоретическая часть 3](#_Toc66742381)

[Нормальное распределение 3](#_Toc66742382)

[Равномерное распределение 3](#_Toc66742383)

[Критерий согласия Пирсона 3](#_Toc66742384)

[Проверка гипотезы 4](#_Toc66742385)

[Квантили распределения 4](#_Toc66742386)

[Правило Стерджиса 5](#_Toc66742387)

[Практическая часть 5](#_Toc66742388)

[Выбранные параметры и язык реализации 5](#_Toc66742389)

[Постановка экспериментов 5](#_Toc66742390)

[Результаты 6](#_Toc66742391)

[Заключение 7](#_Toc66742392)

[Список литературы 7](#_Toc66742393)

# Постановка задачи

1. Сгенерировать выборки из нормального и равномерного распределения
2. Определить при каком количестве измерений возможно различить выборки с помощью критерия согласия Пирсона

# Теоретическая часть

## Нормальное распределение

Нормальное распределение имеет плотность:

В этой формуле µ и σ фиксированные параметры, µ – *среднее значение*, σ – *стандартное отклонение (дисперсия)*.

## Равномерное распределение

Говорят, что случайная величина имеет *непрерывное равномерное распределение* на отрезке , если её плотность имеет вид:

## Критерий согласия Пирсона

Критерий  - статистический критерий для проверки гипотезы , что наблюдаемая случайная величина подчиняется некому теоретическому закону распределения.  
Определение

Пусть дана случайная величина X .

**Гипотеза**: с. в. X подчиняется закону распределения .

Для проверки гипотезы рассмотрим выборку, состоящую из n независимых наблюдений над с.в. X: . По выборке построим эмпирическое распределение  с.в X. Сравнение эмпирического  и теоретического распределения  (предполагаемого в гипотезе) производится с помощью специально подобранной функции — критерия согласия. Рассмотрим критерий согласия Пирсона (критерий ):

**Гипотеза**: порождается функцией  .

Разделим [a,b] на k непересекающихся интервалов ;

Пусть  - количество наблюдений в j-м интервале: ;

 - вероятность попадания наблюдения в j-ый интервал при выполнении гипотезы ;

 - ожидаемое число попаданий в j-ый интервал;

**Статистика:**  – Распределение хи-квадрат с k-1 степенью свободы.

### Проверка гипотезы

В зависимости от значения критерия , гипотеза  может приниматься, либо отвергаться:

* , гипотеза  выполняется.
* (попадает в левый "хвост" распределения). Следовательно, теоретические и практические значения очень близки.   
  Если, к примеру, происходит проверка генератора случайных чисел, который сгенерировал n чисел из отрезка [0,1] и гипотеза : выборка  распределена равномерно на [0,1], тогда генератор нельзя называть случайным (гипотеза случайности не выполняется), т.к. выборка распределена слишком равномерно, но гипотеза  выполняется.
* (попадает в правый "хвост" распределения) гипотеза  отвергается.

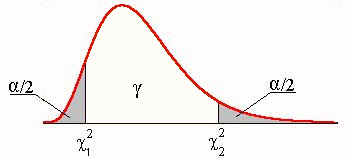


Рисунок 1. Распределение хи-квадрат

### Квантили распределения

**Квантиль хи-квадрат** — это число (величина хи-квадрат), при котором функция распределения хи-квадрат равна заданной (затребованной) вероятности *а*.

Равенство функции распределения хи-квадрат вероятности *а* означает, что с вероятностью *а* будут наблюдаться значения хи-квадрат, не большие, чем найденный (определенный согласно функции распределения) квантиль хи-квадрат. Таким образом, найти квантиль означает разграничить распределения хи-квадрат согласно заданной вероятности *а*.

Существуют готовые таблицы квантилей распределения хи-квадрат, где можно посмотреть необходимые значения.

## Правило Стерджиса

**Правило Стёрджеса** — эмпирическое правило определения оптимального количества интервалов, на которые разбивается наблюдаемый диапазон изменения случайной величины при построении гистограммы плотности её распределения.

Количество интервалов {\displaystyle n}определяется как: ,,, где {\displaystyle N}N — общее число наблюдений величины. {\displaystyle x}

# Практическая часть

## Выбранные параметры и язык реализации

Для выполнения данный работы был выбран язык **R.**

Гипотезы для проверки были сформулированы следующим образом:

* - выборка соответствует нормальному распределению
* – противоположность : выборка не соответствует нормальному распределению

Выборки генерировались разного размера из:

* Нормального распределения:
* Равномерного распределения:

Размеры выборок:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 15 | 20 | 25 | 30 | 40 | 50 | 75 | 100 | 125 | 150 | 175 | 200 | 250 | 300 | 500 | 750 | 1000 |

В зависимости от размера выборки вычислялось кол-во интервалов по критерию Стёрджеса. Все интервалы кроме двух крацних всегда брались равными. Два крайних интервала растягивались так, чтобы объединение всех интервалов покрывало промежуток [-20, 20]. Это было сделано для того, чтобы:

1. все значения сгенерированные по нормальному распределению попадали в интервалы
2. при не самых малых размерах выборки (больше 75 элементов) получалось что в каждом интервале есть хотябы 5 значений, что является желательным для качественной работы критерия хи-квадрат

Также в зависимости от размера выборки выбирался порог принятия гипотезы . Уровень значимости всегда брался равным 0.05. Проверялась всегда только правая граница: можно ли с достаточно большой уверенностью утверждать, что гипотеза не верна.

## Постановка экспериментов

Шаги эксперемента:

1. Генерация выборки заданного размера из нормального/равномерного распределения
2. Вычисление кол-ва значений пришедшихся на каждый интервал
3. Вычисление хи-квадрат статистики

Описанный выше эксперимент проводился по 2000 раз для каждого исследуемого размера выборки (1000 раз для выборок из нормального распределения и 1000 раз для выборок из равномерного распределения). Вычислялось суммарное количество принятий гипотезы для каждого из распределений. Результаты были переведены в процент принятия от всего числа эксперементов для данного размера выборки из конкретного распределения

Замечание: Теоретические значения для вычисления хи-квадрат статистики вычислялись по одному разу для каждого из размеров выборок.

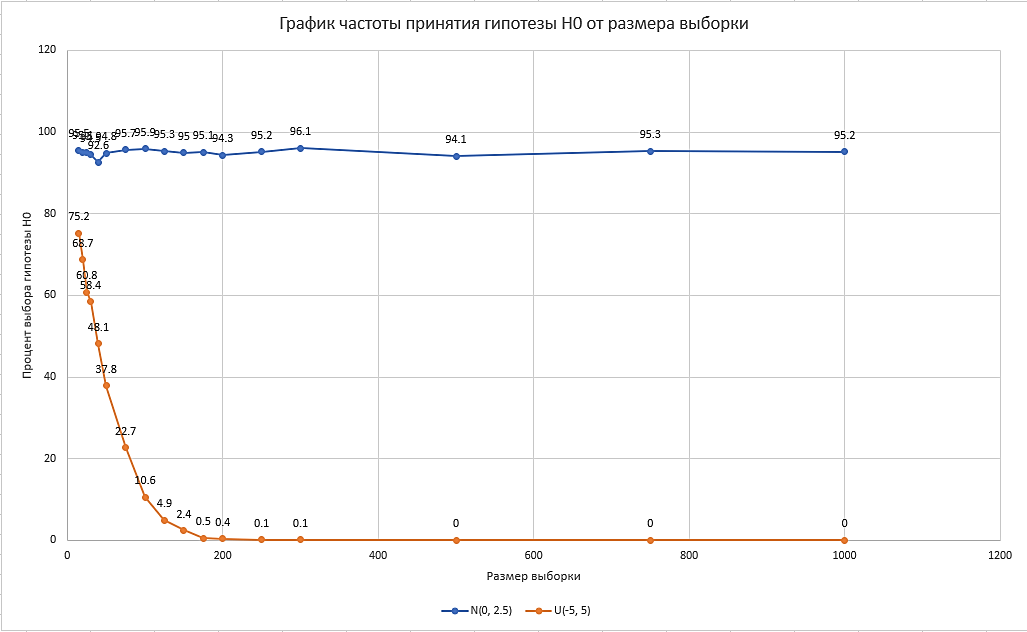
## Результаты

В результате работы была получена таблица вероятностей принятия гипотезы для выборок разного размера из нормального и равномерного распределений (вероятность представлена в виде процента от общего числа экспериментов на выборках из данного распределения)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** | **15** | **20** | **25** | **30** | **40** | **50** | **75** | **100** | **125** |
| **N(0, 2.5)** | **95.5** | **95.1** | **95** | **94.5** | **92.6** | **94.8** | **95.7** | **95.9** | **95.3** |
| **U(-5, 5)** | **75.2** | **68.7** | **60.8** | **58.4** | **48.1** | **37.8** | **22.7** | **10.6** | **4.9** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** | **150** | **175** | **200** | **250** | **300** | **500** | **750** | **1000** |
| **N(0, 2.5)** | **95** | **95.1** | **94.3** | **95.2** | **96.1** | **94.1** | **95.3** | **95.2** |
| **U(-5, 5)** | **2.4** | **0.5** | **0.4** | **0.1** | **0.1** | **0** | **0** | **0** |

Также в ходе работы был построен график визуализирующий значения, представленные в таблице выше



# Заключение

На основе данных полученных в ходе лабораторной можно сделать следующие выводы:

1. Выборки из нормального распределения практически вне зависимости от их размера распознаются как вборки из нормального распределения (при использованном в работе способе разбиения на интервалы)
2. С увеличением размера выборки вероятность верно определить, что выборка сгенерированная по равномерному распределению не является выборкой из нормального распределения существенно возрастает. (При размере выборки большем 200 ошибок практически не бывает)
3. Если использовать выборку размером менее 200 элементов, то с не малой вероятностью эта выборка будет ошибочно распознана как выборка из нормального распределения. Причём незначительное увеличение размера выборки будет вести к существенному увеличению вероятности избежания ошибки (это заключение будет тем вернее, чем меньше был изначальный размер выборки)

Файл кода и этот отчёт можно посмотреть [здесь](https://drive.google.com/drive/folders/1paP1s6V7YHSOTYIQvuBfuynNieqi4ioK?usp=sharing).

# Список литературы

1. Равномерное распределение [Электронный ресурс]   
   url: http://statistica.ru/theory/ravnomernoe-raspredelenie/  
   Дата обращения: 15.03.21
2. Нормальное распределение [Электронный ресурс]   
   url: http://statistica.ru/theory/normalnoe-raspredelenie/  
   Дата обращения: 15.03.21
3. Критерий Пирсона [Электронный ресурс]  
   url: http://www.machinelearning.ru/wiki/index.php?title=Критерий\_хи-квадрат   
   Дата обращения: 15.03.21
4. Квантили хи-квадрат [Электронный ресурс]  
   url: https://ru.wikipedia.org/wiki/Квантили\_распределения\_хи-квадрат   
   Дата обращения: 15.03.21
5. Правило Стёрджеса [Электроный ресурс]  
   url: https://ru.wikipedia.org/wiki/Правило\_Стёрджеса#:~:text=Правило%20Стёрджеса%20—%20эмпирическое%20правило%20определения,Sturges%2C%201882—1958).   
   Дата обращения: 15.03.21