**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра математического обеспечения и применения ЭВМ**

отчет

**по лабораторной работе №4**

**по дисциплине «Статистические методы обработки**

**экспериментальных данных»**

Тема: Элементы корреляционного анализа.

Проверка статистической гипотезы о равенстве

коэффициента корреляции нулю.

|  |  |  |
| --- | --- | --- |
| Студентка гр. 6382 |  | Фиалковский М.С. |
| Преподаватель |  | Середа В.И. |

Санкт-Петербург

2020

**Цель работы.**

Освоение основных понятий, связанных с корреляционной зависимостью между случайными величинами, доверительными интервалами, статистическими гипотезами и проверкой их «справедливости».

**Постановка задачи.**

1. Из заданной генеральной совокупности сформировать выборку по второму признаку. Провести статистическую обработку второй выборки в объеме лабораторных работ №1 и №2, с целью определения точечных статистических оценок параметров распределения исследуемого признака (математического ожидания, дисперсии, среднеквадратического отклонения, асимметрии и эксцесса).

2. Для системы двух случайных величин (первый признак) и (второй признак):

* Сформировать двумерную выборку, построить корреляционную таблицу и найти – значение выборочного коэффициента корреляции. Для расчета использовать формулу:

Где , – количество интервалов в интервальных рядах для выборок значений случайных величин и соответственно; , – условные варианты, значения которых соответствуют выборочным (интервальным) значениям случайных величин и соответственно; – общий объем выборки.

* Построить доверительный интервал для коэффициента корреляции;
* Осуществить проверку статистической гипотезы о равенстве коэффициента корреляции нулю.

**Основные теоретические положения.**

*Коэффициент корреляции Пирсона* характеризует существование линейной зависимости между двумя величинами.

Пусть даны два интервальных ряда и тогда точечная оценка коэффициента корреляции Пирсона рассчитывается по формуле:

Где , – количество интервалов в интервальных рядах для выборок значений случайных величин и соответственно; , – условные варианты, значения которых соответствуют выборочным (интервальным) значениям случайных величин и соответственно; – общий объем выборки.

Можно перейти к условным вариантам, тогда формула приобретает следующий вид:

*Построение доверительного интервала для коэффициента корреляции Пирсона.*

Распределение выборочного коэффициента корреляции Пирсона сложное, поэтому используют преобразование Фишера:

Для полученного среднеквадратическое отклонение равно , где – объем выборки.

Доверительный интервал имеет следующий вид:

Выполнив обратное преобразование, получим доверительный интервал для коэффициента корреляции Пирсона:

где – левая граница доверительного интервала для , а – правая граница доверительного интервала для .

*Проверка гипотезы о равенстве коэффициента корреляции Пирсона нулю.*

Основная и альтернативная гипотезы имеет следующий вид:

Статистика критерия по выборке вычисляется по следующей формуле:

Статистика критерия имеет распределение Стьюдента с степенями свободы. Для заданного уровня значимости по таблице находят критическое значение .

Тогда если , то имеется достаточно оснований, чтобы отклонить гипотезу на заданном уровне значимости. В противном случае, оснований недостаточно.

**Ход работы.**

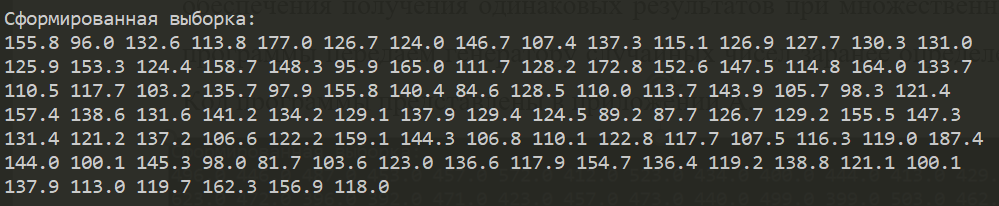
Произведём выборку из представленной генеральной совокупности экспериментальных данных размером 96 элементов. Результат на рис. 1. Для обеспечения получения одинаковых результатов при множественных запусках программы передаём генератору случайных чисел заранее определенное зерно. Код программы представлены в приложении А.

Рис. 1. Начальная выборка

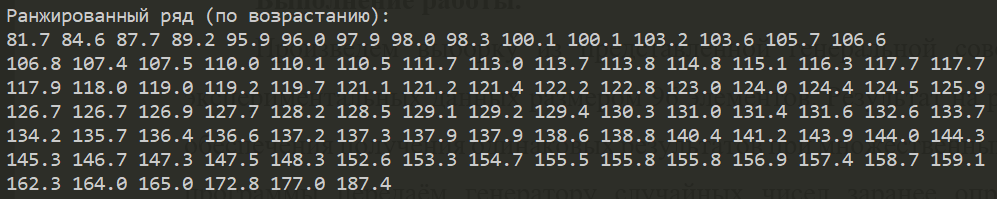
Ранжированный ряд, полученный из выборки, представлен на рис. 2.

Рисунок 2 – Ранжированный ряд.

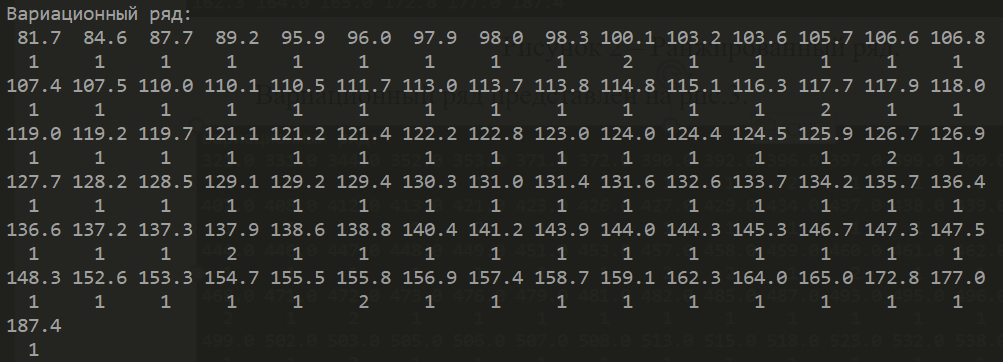
Вариационный ряд представлен на рис.3.

Рис. 3. Вариационный ряд

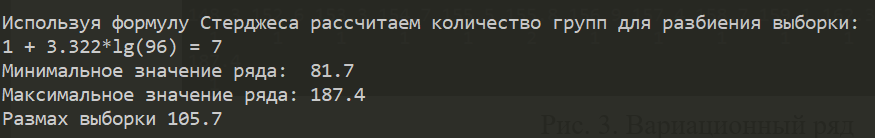
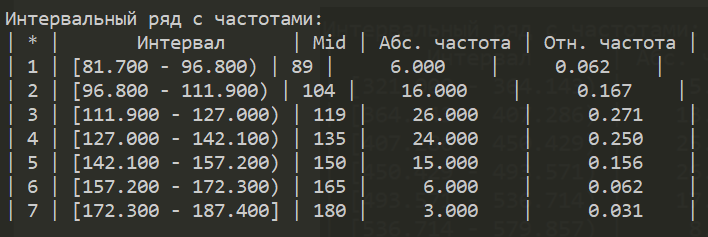
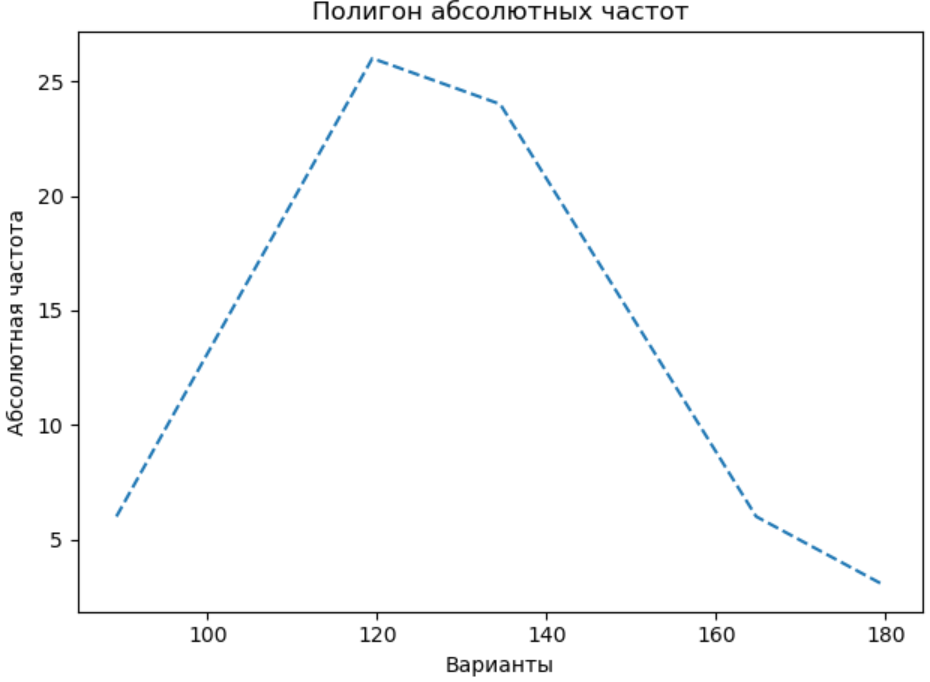
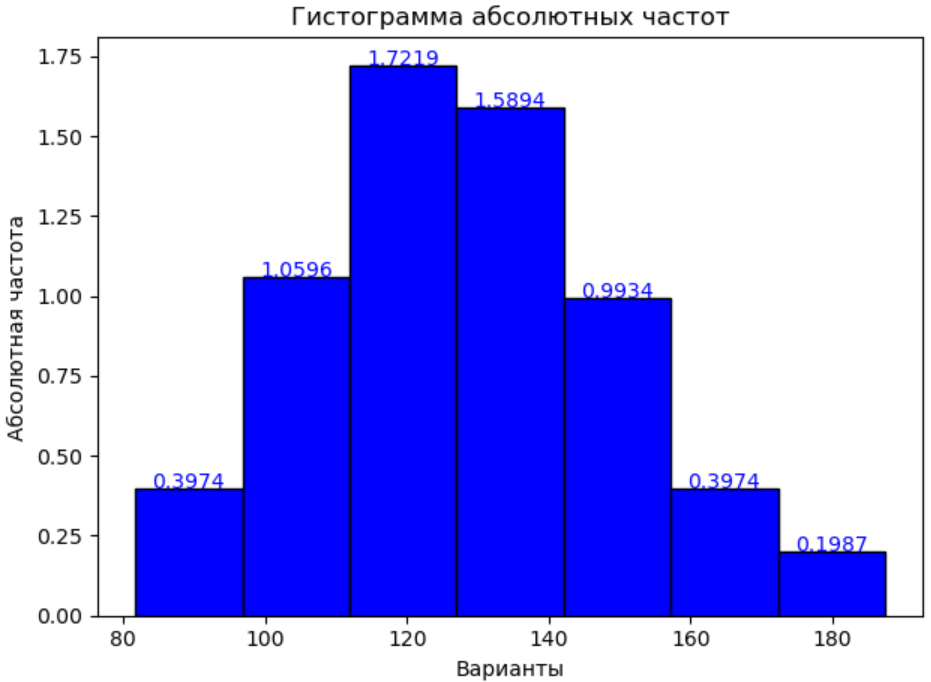
Для построения интервальный ряда требуется рассчитать количество различных групп. Требуемые расчёты и интервальный ряд представлены на рис. 4.

Рис. 4. Расчёты для интервального ряда

Визуализируем полученный интервальный ряд, построив гистограмма и полигон абсолютных (рис. 6) и относительных частот (рис. 7).

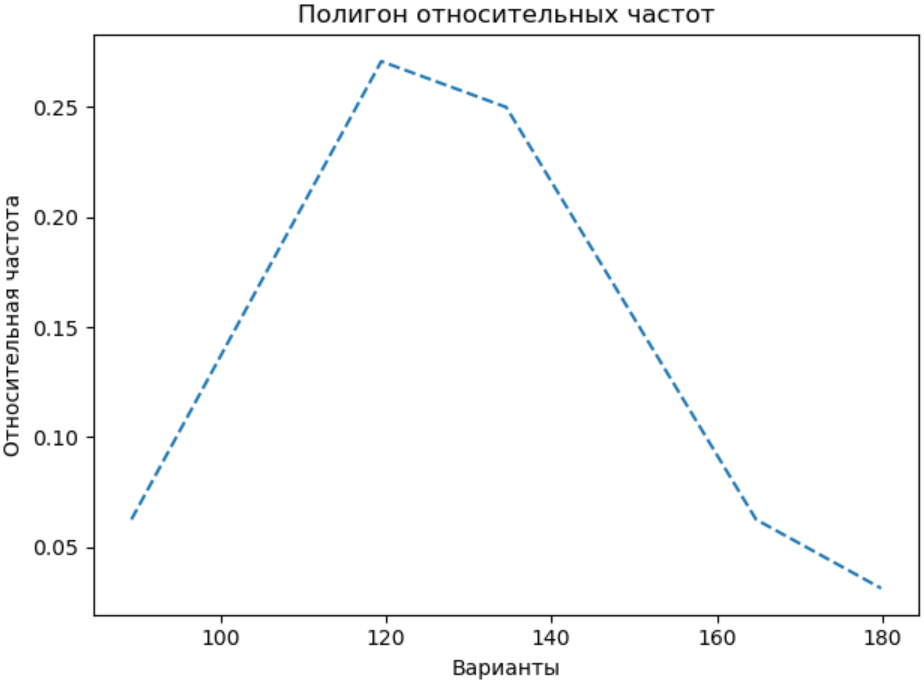
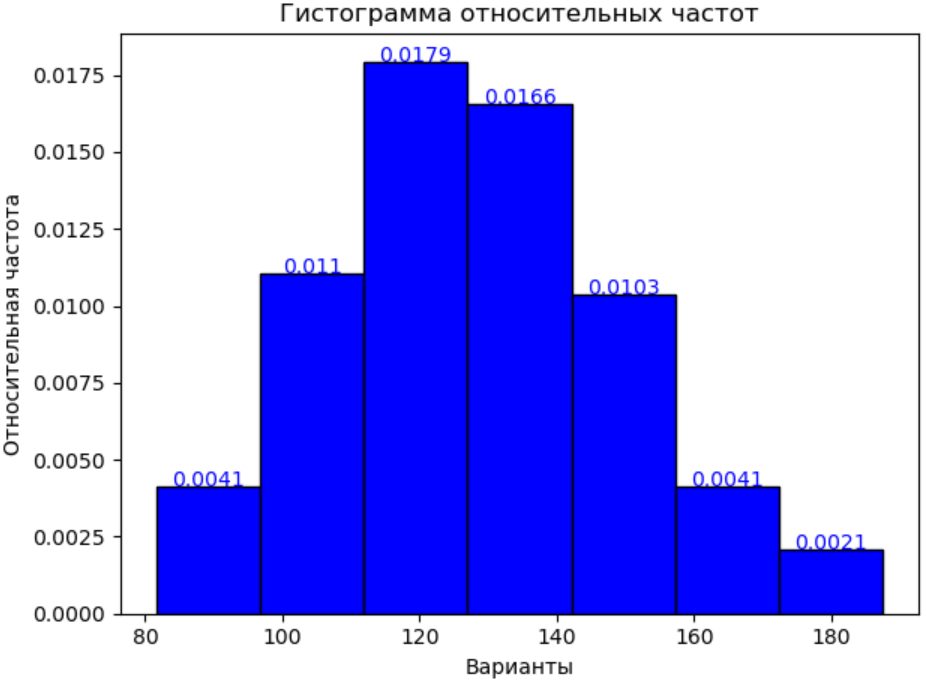
Рис. 5. Гистограмма и полигон абсолютных частот

Рис. 6. Гистограмма и полигон относительных частот

Далее для оценки параметров распределения построим таблицу ищ второй лабораторной работу, используя построенный только что интервальный ряд. Условные варианты вычислены по формуле , где , .

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

Правильность составления таблицы проверим по этой формуле:

Равенство выполняется, поэтому таблица построена верно.

С использованием построенной таблицы найдем точечные оценки параметров распределения исследуемой величины, опустив при этом выкладки вычислений выборочных и центральных выборочных моментов.

* Оценка математического ожидания
* Смещенная оценка дисперсии
* Оценка среднеквадратического отклонения
* Исправленная оценка дисперсии
* Исправленная оценка СКО
* Оценка асимметрии
* Оценка эксцесса

*Оценка коэффициента корреляции.*

Сформируем двумерную выборку. Полученная выборка представлена на рис. 8.

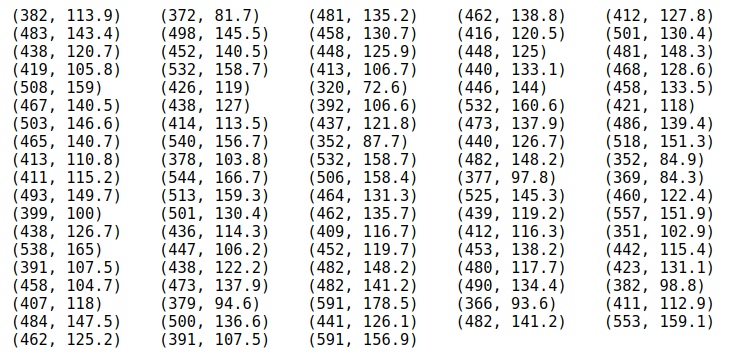


Рисунок 8 – Полученная двумерная выборка.

Построим корреляционную таблицу:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |  | |
|  | |  |  |  |  |  |  |  |

Оценки математического ожидания условных вариант:

Оценки стандартных отклонений уловных вариант:

Тогда оценка коэффициента корреляции:

*Доверительный интервал для коэффициента корреляции.*

В начале построим доверительный интервал для

Где величина находится из следующей формулы с использованием преобразования Фишера:

А величина находится из формулы .

Найдем границы доверительного интервала для .

Тогда получим доверительный интервал:

Чтобы перейти к доверительному интервалу для коэффициента корреляции, воспользуемся обратным преобразованием:

Здесь и – соответственно левая и правая границы доверительного интервала для .

Получим доверительный интервал, покрывающий истинное значение параметра с 95% доверительной вероятностью:

*Проверка статистической гипотезы о равенстве коэффициента корреляции нулю.*

Сформулируем нулевую и альтернативную гипотезы:

: Коэффициент корреляции равен нулю.

: Коэффициент корреляции не равен нулю.

Проверим гипотезу на уровне значимости .

Для этого найдем значение статистики критерия:

Полученное по выборке значение статистики критерия соответствует

Статистика критерия имеет распределение Стьюдента с степенями свободы.

Тогда для заданного уровня значимости критическое значение .

Таким образом, у нас достаточно оснований отвергнуть гипотезу в пользу гипотезы , поскольку .

**Выводы.**

В ходе выполнения лабораторной работы была обработана выборка по второму признаку. Были сформированы ранжированный ряд, вариационный ряд, а также интервальный ряд. По полученному интервальному ряду были построены гистограммы и полигоны абсолютных и относительных частот, а также эмпирическая функция распределения. По интервальному ряду были получены точечные оценки параметров распределения генеральной совокупности с помощью метода моментов.

Далее была сформирована двумерная выборка, по которой построена корреляционная таблица для условных вариант, после чего вычислена точечная оценка коэффициента корреляции Пирсона. Полученная оценка коэффициента корреляции равна , что говорит о наличии корреляционной зависимости между исследуемыми величинами.

После был построен доверительный интервал для коэффициента корреляции Пирсона. Полученный доверительный интервал уровня доверия имеет следующий вид: . Данный доверительный интервал покрывает истинное значение коэффициента корреляции с доверительной вероятностью 95%.

Далее была проверена гипотеза о равенстве коэффициента корреляции нулю. В результате было получено выборочное значение статистики критерия , а критическое значение для уровня значимости равно . Таким образом, имеется достаточно оснований, чтобы отклонить нулевую гипотезу о том, что коэффициент корреляции равен нулю. Полученный результат указывает на то, что вполне вероятно, что корреляционная зависимость между исследуемыми параметрами действительно существует, а не была получена случайно.

**ПРИЛОЖЕНИЕ А**

**ИСХОДНЫЙ КОД**

import pandas as pd

from collections import defaultdict

from math import log, floor

import matplotlib.pyplot as plt

import numpy as np

from math import sqrt, log, exp

from scipy.stats import t

get\_ipython().run\_line\_magic('matplotlib', 'inline')

from IPython.core.pylabtools import figsize

figsize(7, 7)

# чтение данных

df = pd.read\_excel('./data.xlsx', head=True)

# формируем случайную выборку заданного размера

n = 94

x = df.sample(n, replace=True, random\_state=5)

x = x.reset\_index(drop=True)

# выборки

val\_x = list(x['v'])

val\_y = list(x['E'])

# объем выборки

n = 94

# границы интервалов

bins\_x = [320]

h = 39 # шаг

for i in range(1, 8):

bins\_x.append(bins\_x[i-1] + h)

bins\_y = [72.6]

h = 15.2 # шаг

for i in range(1, 8):

bins\_y.append(bins\_y[i-1] + h)

# частоты

vals\_x = [4, 11, 16, 30, 21, 9, 3]

vals\_y = [5, 6, 20, 23, 25, 12, 3]

# построение корреляционной таблицы

def check\_interval(x, y, x\_int, y\_int):

first = (x >= x\_int[0] and x < x\_int[1])

second = (y >= y\_int[0] and y < y\_int[1])

return first and second

sample = list(zip(val\_x, val\_y))

table = []

for i in range(1, 8):

a = (bins\_x[i-1], bins\_x[i])

table.append([])

for j in range(1, 8):

b = (bins\_y[j-1], bins\_y[j])

tmp = map(lambda x: check\_interval(x[0], x[1], a, b), sample)

table[i-1].append(sum(tmp))

table

for i in range(1, len(sample)):

print(sample[i-1], end='\t')

if i % 5 == 0:

print()

# шаги интервальных рядов

h\_x = bins\_x[1] - bins\_x[0]

h\_y = bins\_y[1] - bins\_y[0]

# середины интервалов в интервальных рядах

med\_x = [339, 378, 417, 456, 495, 534, 573]

med\_y = [80.2, 95.4, 110.6, 125.8, 141.0, 156.2, 171.4]

# условные варианты

v = list(map(lambda x: (x - 456)/h\_x, med\_x))

u = list(map(lambda x: (x - 125.8)/h\_y, med\_y))

# считаем сумму из формулы

s = 0

for i in range(7):

for j in range(7):

s += table[i][j] \* v[i] \* u[j]

# средние и СКО для условных вариант

mean\_v = sum(map(lambda x: x[0] \* x[1], zip(v, vals\_x))) / n

M2\_v = sum(map(lambda x: x[0]\*\*2 \* x[1], zip(v, vals\_x))) / n

m2\_v = M2\_v - mean\_v\*\*2

sd\_v = sqrt(m2\_v)

mean\_u = sum(map(lambda x: x[0] \* x[1], zip(u, vals\_y))) / n

M2\_u = sum(map(lambda x: x[0]\*\*2 \* x[1], zip(u, vals\_y))) / n

m2\_u = M2\_u - mean\_u\*\*2

sd\_u = sqrt(m2\_u)

print('v: mean = {}, sd = {}'.format(mean\_v, sd\_v))

print('u: mean = {}, sd = {}'.format(mean\_u, sd\_u))

# коэффициент корреляции Пирсона

r = (s + n \* mean\_v \* mean\_u)/(n \* sd\_v \* sd\_u)

print('Коэффициент корреляции Пирсона {}'.format(r))

# проверка гипотезы о равенстве коэф корреляции нулю

t\_sample = r \* sqrt(n-2) / sqrt(1 - r\*\*2)

t\_cr = t.ppf(0.95, df=n-2)

print('t критическое = {}'.format(t\_cr))

print('t выборочное = {}'.format(t\_sample))

if t\_sample <= t\_cr:

print('Не достаточно отснований, чтобы отклонить H0')

else:

print('Отвергаем гипотезу H0')

# нахождение доверительного интервала

# преобразование Фишера

z = 0.5 \* log((1 + r)/(1 - r))

# доверительный интервал для z

left = z - 1.96 / sqrt(n - 3)

right = z + 1.96 / sqrt(n - 3)

print('Доверительный интервал: ({}, {})'.format(left, right))

# обратное преобразование

left = (exp(2 \* left) - 1) / (exp(2 \* left) + 1)

right = (exp(2 \* right) - 1) / (exp(2 \* right) + 1)

print('Доверительный интервал: ({}, {})'.format(left, right))

Список литературы

1. Лекции по математической статистике Computer Science Center.
2. Лагутин М. Б. Наглядная математическая статистика. (Том 2, стр. 174) — М.: П-центр, 2003.
3. Севастьянов Б. А. Курс теории вероятностей и математической статистики. — Москва-Ижевск: Институт компьютерных исследований, 2004.