МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра систем сбора и обработки данных

КУРСОВОЙ ПРОЕКТ

по дисциплине: Компьютерные технологии моделирования и анализа данных

на тему: Проверка гипотезы о виде распределения

Вариант №26

Факультет: ФПМИ

Группа: ПММ-21

Выполнил: Сухих А.С.

Проверил: д.т.н., профессор Лемешко Борис Юрьевич

Дата выполнения: 09.01.23

Отметка о защите:

Новосибирск 2023

Введение

Цель работы: 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.

Наиболее распространенным законом распределения данных, наблюдаемых в реальном мире, является закон нормального распределения. Принадлежность выборки нормальному закону часто может являться требованием для применения различных методов математической статистики. Определить эту принадлежность можно с помощью проверки на нормальность с помощью различных критериев. В данной работе будет исследовано применение критерия Дэвида–Хартли–Пирсона, основанного на отношении размаха выборки к её выборочному стандартному отклонению.

Постановка задачи

Согласно варианту №26 необходимо выполнить проверку гипотезы о нормальности распределения по критерию Дэвида–Хартли–Пирсона. В данном критерии рассматривается отношение размаха выборки к выборочному стандартному отклонению и его статистика имеет вид

$$U = \frac{R}{s},$$

где $R = x_{max} - x_{min}$ — размах выборки,

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ — несмещенная оценка дисперсии

Критерий двусторонний: гипотеза о нормальности распределения отвергается, если $U < U_{\alpha/2}$ или $U > U_{1-\alpha/2}$. Вывод о справедливости гипотезы

можно сделать, сравнив вычисленную статистику с таблицей процентных точек, представленной в Приложении 1.

Также необходимо вычислить достигаемый уровень значимости статистики критерия. Поскольку для данного критерия неизвестна предельная статистика достигаемый уровень значимости будет вычислен методом Монте-Карло:

1. Вычислить $S = S(X_n)$ — статистику критерия по выборке.
2. Установить $m = 0$.
3. Сгенерировать выборку Y_n при верной гипотезе H_0.
4. Вычислить значения $S(Y_n)$.
5. Если $S(Y_n) > S(X_n)$, то $m = m+1$.
6. Повторять шаги 3-5 N раз.

Оценка достигаемого уровня значимости $\check{p} = 2 \cdot min\left(\frac{m}{N}, 1 - \frac{m}{N}\right)$

Количество повторений N определяется из выражения

$$N = \left[t_\gamma^2 \frac{p(1-p)}{\varepsilon^2} \right] + 1,$$

где γ — доверительная вероятность, t_γ — квантиль стандартного нормального распределения, ε — погрешность моделирования, p — вероятность попадания в доверительную область.

Аналитический обзор

Критерий Дэвида-Хартли-Пирсона был предложен в 1954 году как результат совместного исследования. На момент исследования данного критерия уже были представлены несколько исследований касаемо стьюдентизированного диапазона являющегося отношением размаха выборки размера n генеральной совокупности со стандартным отклонением σ к независимой среднеквадратичной оценке σ. Также уже был представлен критерий Гири (1933).

Однако в исследовании нового критерия фокус был направлен на изучение иной статистики, связанной с отношением размаха выборки к стандартному отклонению, при этом обе величины отношения вычислялись по одной и той же выборке из *n* наблюдений. Так, ранее исследованное отношение могло быть использовано для быстрого анализа отклонения, а статистика нового критерия, которая зависела только от конкретной выборки, позволяла выявлять неоднородность данных и отклонение от нормального распределения.

Предпосылкой к появлению данного критерия стала переписка между одним из исследователей критерия и доктором Джозефом Берксоном, проводившим рутинную проверку данных, сравнивая размах и оценки стандартного отклонения σ. В результате переписки начались эмпирические исследования связи между оценками для определения стандартной ошибки разницы между ними. Также за несколько лет до этого в 1946 году Г. А. Бейкер показал возможность использования такого отношения как критерий однородности, продемонстрировав на эксперименте с искусственной выборкой, что это соотношение будет значительно зависеть от формы родительской совокупности.

Результаты исследований

С помощью языка программирования Python было разработано программное обеспечение для проверки гипотезы о соответствии нормальному распределению с применением критерия Дэвида-Хартли-Пирсона. Программное обеспечение представляет собой консольное приложение с возможностью генерации выборки согласно нормальному закону распределения с заданным объемом выборки и параметрами сдвига и масштаба и возможностью проверки выборки на нормальность.

Исследования были проведены на примере выборок, смоделированных в соответствии с нормальным законом, а также на примере реальных данных с различными объемами выборок.

Во всех исследованиях ошибка первого рода α задаётся равной 0,05, а количество повторений в методе Монте-Карло равно 16600, что обеспечивает относительную погрешность моделирования 0,01.

1. Проверка гипотезы по выборкам, смоделированным в соответствии с нормальным законом распределения.

Результаты тестирования программы на смоделированных выборках показаны в таблице 1. Были смоделированы выборки:

- стандартное нормального распределения (сдвиг 0, масштаб 1)
- нормальное со сдвигом 3 и масштабом 1
- нормальное со сдвигом 0 и масштабом 5

Каждая выборка представлена объемом n=100 и n=1000. Проверяется гипотеза о соответствии стандартному нормальному распределению.

Таблица 1. Результаты проверки нормальности смоделированных выборок

Выборка	Статистика U	p-value	Результат
Нормальное N(0,1), n=100	5.04374	0.90651	не отклоняется
N(3,1), n=100	5.10267	0.80578	не отклоняется
N(0,5), n=100	4.92366	0.89819	не отклоняется
N(0,1), n=1000	6.55909	0.79723	не отклоняется
N(3,1), n=1000	6.24	0.65783	не отклоняется
N(0,5), n=1000	6.86053	0.40747	не отклоняется

Проверить справедливость гипотезы также можно с помощью таблицы процентных точек (приложение 1). Для n=100 при вероятности ошибки первого рода 0,05 достоверной областью является интервал значений (4,206; 6,112). Как видно в таблице 1 все значения для n=100 попадают в данный интервал и, следовательно, гипотеза не отклоняется.

По приведенным выше результатам видно, что критерий для всех выборок не отверг гипотезу о нормальности выборки, что верно, так как все

выборки сгенерированы согласно нормальному закону. Однако критерий не смог отличить выборки с измененными параметрами сдвига или масштаба от стандартного нормального распределения даже на сравнительно больших (n=1000) объемах выборки.

Данный критерий присутствует также в программе ISW. Результаты проверки тех же выборок при α=0,05 и количестве повторений N=16600 на нормальность в ISW приведены в таблице 2.

Таблица 2. Результаты проверки нормальности смоделированных выборок в ISW

Выборка	Статистика U	p-value	Результат
Нормальное N(0,1), n=100	5.04374	0.903735	не отклоняется
N(3,1), n=100	5.10267	0.81145	не отклоняется
N(0,5), n=100	4.92366	0.903735	не отклоняется
N(0,1), n=1000	6.55909	0.79145	не отклоняется
N(3,1), n=1000	6.24	0.638	не отклоняется
N(0,5), n=1000	6.86053	0.402	не отклоняется

При сравнении результатов из таблиц 1 и 2 можно увидеть, что значения статистик для всех выборок в точности совпадают. Однако разница в значениях достигаемого уровня значимости может отличаться на несколько процентов, что связано с фактором случайности при моделировании методом Монте-Карло.

При повторной проверке гипотезы в разработанной программе значение достигаемого уровня значимости также может изменяться на несколько процентов, в то время как в ISW p-value остаётся прежним. Это связано с неизменным начальным значением генератора случайных чисел в ISW, из-за чего моделирование методом Монте-Карло производится с применением одних и тех же случайных выборок.

2. Проверка гипотезы на реальных данных.

В качестве примеров реальных данных были взяты несколько выборок возраста, веса, систолического давления и заработной платы.

Выборка X_1 веса из опроса 40 женщин и 40 мужчин [5]:

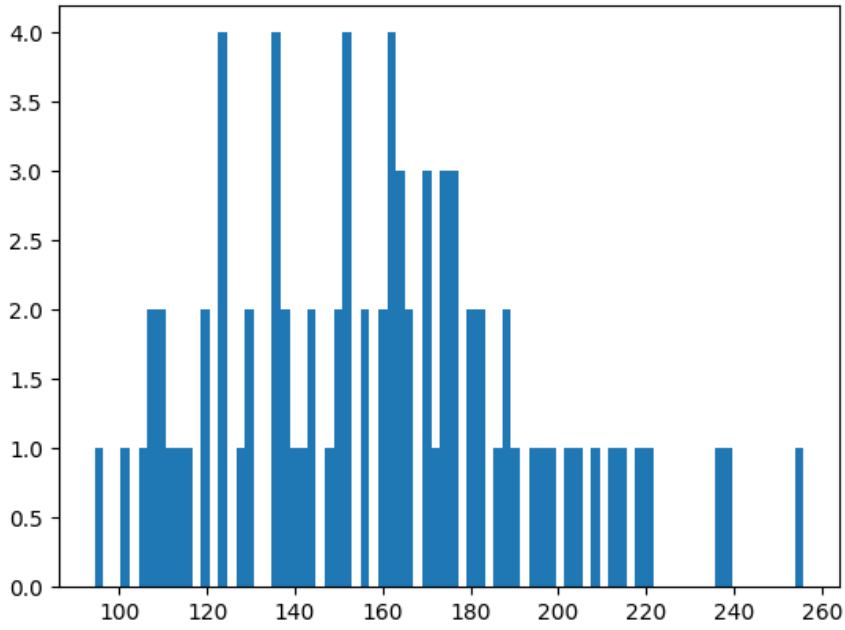


Рисунок 1. Гистограмма частот выборки веса (фунты)

Выборка X_2 систолического давления из опроса 40 женщин и 40 мужчин [5]:

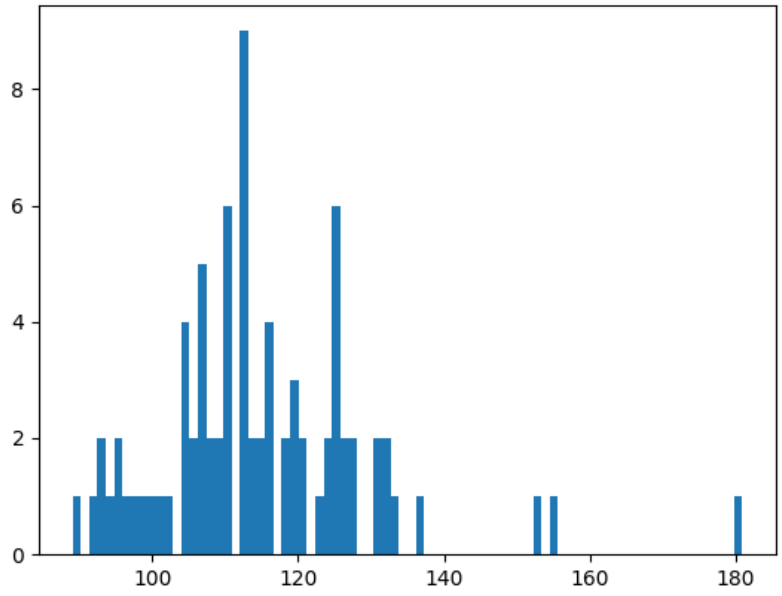


Рисунок 2. Гистограмма частот выборки систолического давления (мм рт.ст.)

Выборка X_3 возраста из опроса 40 женщин и 40 мужчин [5]:

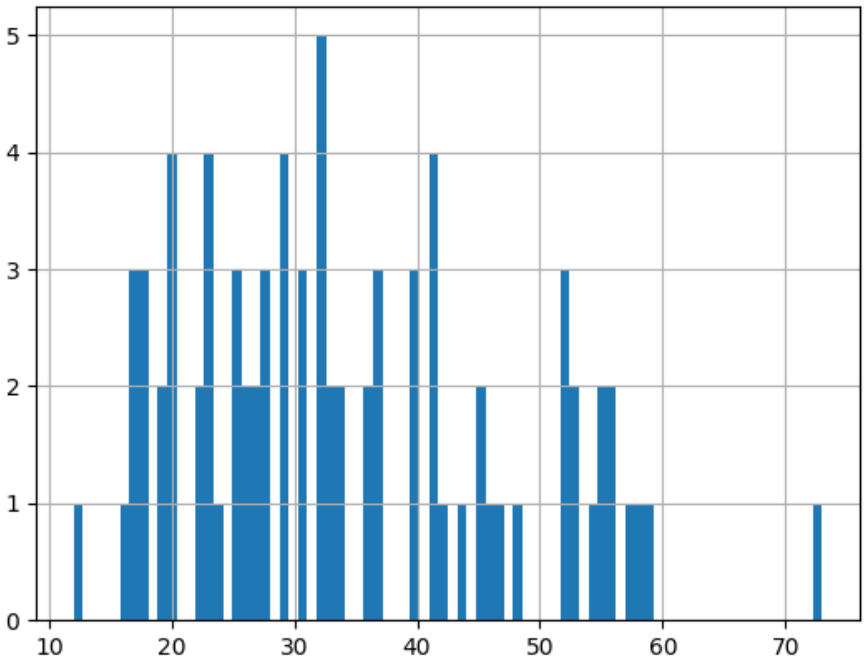
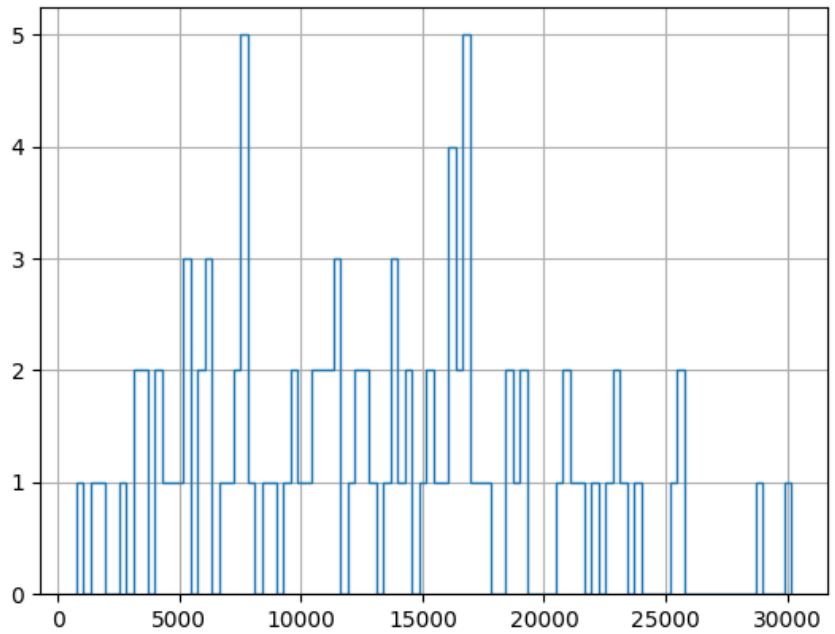


Рисунок 3. Гистограмма частот выборки возраста (лет)

Выборки $X_{4\text{-}6}$ начисление заработной платы в округе Кук, штат Иллинойс в первом квартале 2018 года [6] с объемами выборок n=100, 1000, 26539.



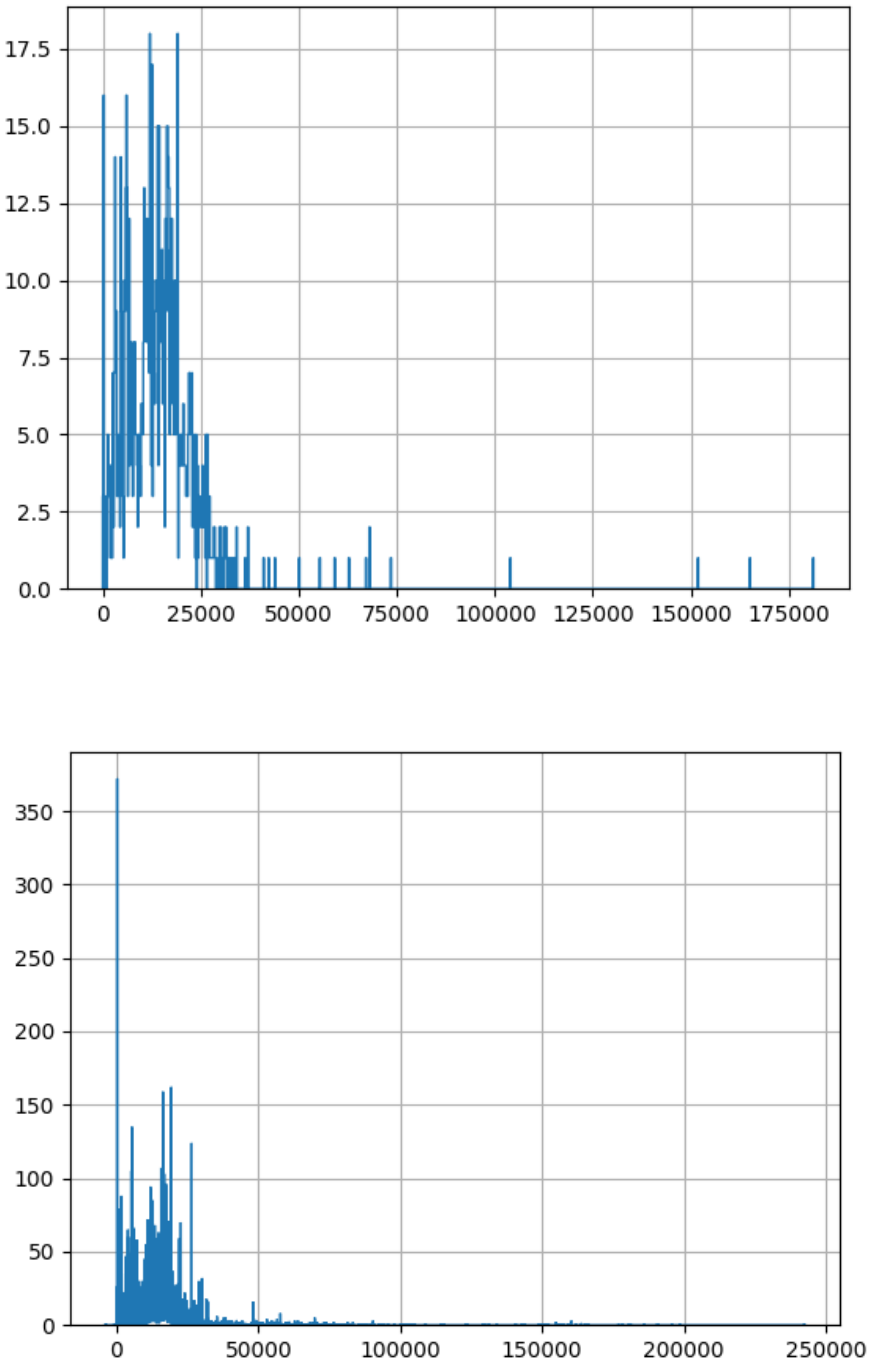


Рисунок 4. Гистограмма частот выборки зарплаты ($)

Результаты проверок этих выборок на нормальность приведены в таблице 3.

Таблица 3. Результаты проверки нормальности реальных данных

Выборка	Статистика U	p-value	Результат
X_1 (вес), n=80	4.63344	0.68723	не отклоняется
X_2 (давление), n=80	6.2708	0.01313	отклоняется

X_3 (возраст), n=80	4.62976	0.67048	не отклоняется
X_4 (зарплата), n=100	4.39404	0.16193	не отклоняется
X_5 (зарплата), n=1000	14.69406	0	отклоняется
X_6 (зарплата), n=26539	21.50086	0	отклоняется

Как можно увидеть по результатам существенный вклад в вывод о справедливости гипотезы вносят аномальные наблюдения в выборках. Так в выборках X_1-X_4 аномальных наблюдений достаточно мало и их значение отличается не более чем в два раза от среднего, вследствие чего гипотеза о нормальности не отклоняется. В выборках X_5-X_6 присутствуют аномальные наблюдения справа и их значения превышают среднее в десятки и сотни раз. Это приводит к большому значению статистики и достигаемый уровень значимости становится равным нулю, гипотеза отклоняется.

Выводы

В результате работы над курсовым проектом было разработано программное обеспечение для проверки гипотезы о соответствии нормальному распределению с применением критерия Дэвида-Хартли-Пирсона.

Программа была протестирована на выборках, сгенерированных в соответствии с законом нормального распределения. Для всех выборок гипотеза не была отклонена. Было выполнено сравнение результатов работы программы с результатами проверки критерием в ISW. Статистики критериев точно совпали, но значения достигаемых уровней значимости различались в пределах нескольких процентов.

По результатам проверки смоделированных выборок было определено, что критерий Дэвида-Хартли-Пирсона способен определить нормальность выборки, но не способен выявить отклонение от нормального закона изменением параметра сдвига или масштаба на малых объемах выборки.

Также были выполнены проверки на нормальность выборок реальных данных: веса, систолического давления, возраста и зарплат. По полученным результатам можно сделать вывод о существенном влиянии на результат аномальных наблюдений, значительно удаленных от среднего арифметического. Для более эффективной проверки критерием Дэвида-Хартли-Пирсона целесообразно применить цензурирование или отбраковку аномальных наблюдений.

Критерий Дэвида-Хартли-Пирсона показал бóльшую достоверность на больших объемах выборок, что подтверждается исследованием мощности критерия в [3]. Данный критерий применим для проверки на нормальность и однородность выборок, однако его применение на малых объемах выборок нецелесообразно, а для больших объемов существуют более мощные критерии.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. David H. A., Hartley H. O., Pearson E. S. The distribution of the ratio, in a single normal sample, of range to standard deviation //Biometrika. – 1954. – Т. 41. – №. 3/4. – С. 482-493.

2. Кобзарь А. И. Прикладная математическая статистика. – 2006.

3. Лемешко Б.Ю., Рогожников А.П. Исследование особенностей и мощности некоторых критериев нормальности // Метрология. 2009. № 4. – С. 3-24.

4. Лемешко Б. Ю., Постовалов С. Н., Лемешко С. Б. Компьютерные технологии анализа данных и исследования статистических законномерностей. – 2007.

5. Statistics Data Sets [Электронный ресурс] // Matt Teachout. College of the Canyons Math Department. URL: http://www.matt-teachout.org/data-sets-for-stats.html. (дата обращения: 01.09.2023)

6. Employee Payroll [Электронный ресурс] // Data.gov. URL: https://catalog.data.gov/dataset/employee-payroll. (дата обращения: 01.09.2023)

ПРИЛОЖЕНИЯ

Приложение 1. Таблица процентных точек статистики критерия Дэвида-Хартли-Пирсона для некоторых объемов выборки.

n	0.15		0.1		0.05		0.025		0.01	
	$\alpha/2$	$1-\alpha/2$	$\alpha/2$	$1-\alpha/2$	$\alpha/2$	$1-\alpha/2$	$\alpha/2$	$1-\alpha/2$	$\alpha/2$	$1-\alpha/2$
10	2.723	3.624	2.670	3.686	2.593	3.778	2.530	3.854	2.458	3.936
20	3.240	4.392	3.178	4.488	3.087	4.633	3.012	4.763	2.927	4.915
30	3.535	4.787	3.469	4.896	3.374	5.066	3.293	5.217	3.203	5.400
40	3.741	5.046	3.674	5.162	3.574	5.345	3.493	5.507	3.401	5.708
50	3.900	5.236	3.831	5.356	3.729	5.546	3.644	5.720	3.550	5.929
60	4.028	5.384	3.958	5.508	3.856	5.704	3.769	5.886	3.674	6.106
80	4.230	5.607	4.158	5.735	4.054	5.937	3.967	6.124	3.870	6.354
100	4.382	5.774	4.311	5.905	4.206	6.112	4.117	6.302	4.018	6.536
150	4.656	6.059	4.583	6.191	4.477	6.405	4.388	6.600	4.288	6.838
200	4.847	6.255	4.774	6.388	4.668	6.600	4.578	6.799	4.474	7.044
300	5.111	6.512	5.037	6.645	4.931	6.858	4.841	7.056	4.741	7.303

Приложение 2. Исходный код файла генерации выборок sample_generator.py

```
import random


def gen_sample(capacity, mu=0, sigma=1):
    return [random.normalvariate(mu=mu, sigma=sigma) for _ in range(0, capacity)]


def sample_to_file(sample, mu, sigma, path):
    with open(path, 'w', encoding='cp1251') as f:
        n = len(sample)
        f.write(f"Нормальное распределение с масштабом {sigma} и сдвигом {mu}\n")
        f.write(f"0 {n}\n")
        for x in sample:
            f.write(f"{x}\n")


if __name__ == '__main__':
    print("Создание нормально-распределенной выборки ...\nВведите следующие
параметры.")
    try:
        capacity = int(input("Объем выборки: "))
        mu = int(input("Сдвиг: "))
        sigma = int(input("Масштаб: "))
        sample = gen_sample(capacity, mu, sigma)
        filename = input("Выходной файл: ")
```

```
        sample_to_file(sample, mu, sigma, filename)
    except Exception as e:
        print("Some error", e.__str__())
```

Приложение 3. Исходный код файла проверки выборок на нормальность критерием Дэвида-Хартли-Пирсона main.py

```
      import math
import numpy
from matplotlib import pyplot
from sample_generator import gen_sample


def calculate_criteria_stat(sample):
    sample_range = max(sample) - min(sample)
    sum = 0
    mean = numpy.mean(sample)
    for x in sample:
        sum += (x - mean) ** 2
    dispersion_estimate = 1 / (len(sample) - 1) * sum
    return sample_range / math.sqrt(dispersion_estimate)


def calculate_p_value(sample, iterations, mu=0, sigma=1):
    m = 0
    n = len(sample)
    x_stat = calculate_criteria_stat(sample)
    print(f"X-статистика критерия: {x_stat}")
    for i in range(0, iterations):
        y_sample = gen_sample(n, mu=mu, sigma=sigma)
        y_stat = calculate_criteria_stat(y_sample)
        if y_stat > x_stat:
            m += 1
    if m / iterations < (1 - m / iterations):
        return 2 * m / iterations
    else:
        return 2 * (1 - m / iterations)


def sample_from_file(path):
    with open(path, 'r', encoding='cp1251') as f:
        name = f.readline()
        n = int(f.readline().split(' ')[1])
        sample = list()
        for i in range(0, n):
            sample.append(float(f.readline()))
```

```
    return sample


if __name__ == "__main__":
    print("Проверка нормальности выборки критерием Дэвида-Хартли-Пирсона.")
    filename = input("Файл с выборкой: ")
    mu = int(input("Сдвиг: "))
    sigma = int(input("Масштаб: "))
    alfa = float(input("Вероятность ошибки первого рода: "))
    iterations = int(input("Количество повторений в методе Монте-Карло: "))
    x_sample = sample_from_file(filename)
    pyplot.hist(x_sample, bins=len(x_sample), histtype='step')
    pyplot.grid()
    pyplot.savefig("sample.png")
    p_value = calculate_p_value(x_sample, iterations, mu, sigma)
    print("p-value:", p_value)
    if p_value < alfa:
        print("Достигаемый уровень значимости менее вероятности ошибки первого рода.
Гипотеза ОТКЛОНЯЕТСЯ.")
    else:
        print("Достигаемый уровень значимости более вероятности ошибки первого рода.
Гипотеза НЕ ОТКЛОНЯЕТСЯ.")
```