



Министерство науки и высшего образования Российской Федерации
Калужский филиал
федерального государственного бюджетного
образовательного учреждения высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(КФ МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУК «Информатика и управление»

КАФЕДРА ИУК4 «Программное обеспечение ЭВМ,

информационные технологии»

Домашняя работа №2

«Проверка гипотез»

ДИСЦИПЛИНА: «Методы обработки информации»

Выполнил: студент гр. ИУК4-72Б _____ (_____)
(подпись) (Ф.И.О.)

Проверил: _____ (_____)
(подпись) (Ф.И.О.)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:

- Оценка:

Калуга, 2023

Цель: формирование у студентов практических навыков выбора подходящих методов обработки информации, исходя из требований технического задания; генерации различных вариантов решений поставленных задач и выбора наиболее рациональных вариантов; обосновать принятые решения.

Задачи: обработать выборку с неизвестными параметрами, проверить гипотезу о параметрах распределения, законе распределения используя различные методы и подходы.

Задание 1

Постановка задачи

Пусть проверяется простая гипотеза относительно параметра распределения $H_0: \theta = \theta_0$, с заданным уровнем значимости α .

Для нескольких альтернативных гипотез $H_1: \theta = \theta_{1i}$, при $\theta_{1i} = \theta_0 + i\Delta (i = 1, 2, 3, 4, 5)$. Построить графики мощности критерия значимости, если используется выборка (выборка из ПЗ-2):

1. объема $k_1 = 25$ (любые 25 значений из заданной выборки);
2. объема $k_2 = N$ (полный объем исходной выборки) Используя полученные результаты, построить таблицы "Ошибка II рода и мощность для нескольких альтернативных гипотез с объемом выборки k_i и α " и графики функций мощности критерия для случая 1 и 2.

Листинг программы

```
import argparse
import csv

import matplotlib.pyplot as plt
import numpy as np
from prettytable import PrettyTable
from scipy.stats import norm

def get_test_power(
    data: list[float], theta_0: float, alpha: float, theta: float,
    sample_size: int
) -> float:
    critical_value = norm.ppf(1 - alpha)
    standard_error = np.std(data) / np.sqrt(sample_size)
    critical_region = (
        theta_0 - critical_value * standard_error,
        theta_0 + critical_value * standard_error
    )
```

```

power = 1 - norm.cdf(critical_region[1], theta, standard_error) + norm.cdf(
    critical_region[0], theta, standard_error
)
return power

def read_points(path: str) -> list[float]:
    points = []

    with open(path, newline='') as file:
        reader = csv.reader(file, delimiter=' ', quotechar='|')
        for row in reader:
            points.append(float(''.join(row)))

    return points

def solve(points: list[float]):
    alpha = 0.1
    theta_0 = np.mean(points)
    delta = np.std(points) / np.sqrt(len(points))
    thetas = [theta_0 + i * delta for i in range(1, 6)]

    small_sample_power = np.array(
        [
            get_test_power(
                points, theta_0, alpha, theta, 25
            ) for theta in thetas
        ]
    )
    full_sample_power = np.array(
        [
            get_test_power(
                points, theta_0, alpha, theta, len(points)
            ) for theta in thetas
        ]
    )

    table = PrettyTable()
    table.add_column('Значение параметра распределения', thetas)
    table.add_column('Мощность критерия', small_sample_power)
    table.add_column('Ошибка II рода', 1 - small_sample_power)
    print('Любые 25 из заданной выборки:')
    print(table)

    table = PrettyTable()
    table.add_column('Значение параметра распределения', thetas)
    table.add_column('Мощность критерия', full_sample_power)
    table.add_column('Ошибка II рода', 1 - full_sample_power)
    print('Все значения выборки:')
    print(table)

    plt.figure(figsize=(10, 5))
    plt.plot(thetas, small_sample_power, label='$Мощность критерия при N=25$')
    plt.plot(
        thetas, full_sample_power, label=f'Мощность критерия $N={len(points)}$'
    )
    plt.xlabel('Значение параметра распределения')
    plt.ylabel('Мощность критерия')
    plt.title('График мощности критерия')
    plt.axhline(
        alpha, color='red',

```

```

        label=f"$\\alpha={alpha}$"
    )
    plt.legend()
    plt.show()

if __name__ == '__main__':
    parser = argparse.ArgumentParser()
    parser.add_argument('-file')

    args = parser.parse_args()
    file = args.file or './data/Test14.csv'

    points = read_points(file)
    solve(points)

```

Результат выполнения задания

За нулевую гипотезу будем брать среднее значение выборки.

Уровень значимости $\alpha = 0.1$.

Все значения выборки:

Значение параметра распределения	Мощность критерия	Ошибка II рода
0.1461757984121978	0.4004016061579658	0.5995983938420342
0.22031553737164683	0.7642757723355271	0.23572422766447287
0.2944552763310958	0.9571518535622745	0.0428481464377255
0.36859501529054484	0.9967206201316634	0.003279379868336596
0.44273475424999387	0.999899775068096	0.00010022493190398052

Рисунок 1 – Ошибки II рода и мощность критерия для полной выборки

Любые 25 из заданной выборки:

Значение параметра распределения	Мощность критерия	Ошибка II рода
0.1461757984121978	0.24078266609083318	0.7592173339091668
0.22031553737164683	0.35300632649265795	0.646993673507342
0.2944552763310958	0.508814786124697	0.491185213875303
0.36859501529054484	0.6712869326824206	0.32871306731757943
0.44273475424999387	0.8081769602965775	0.19182303970342252

Рисунок 2 – Ошибки II рода и мощность критерия для малой выборки

($N = 25$)

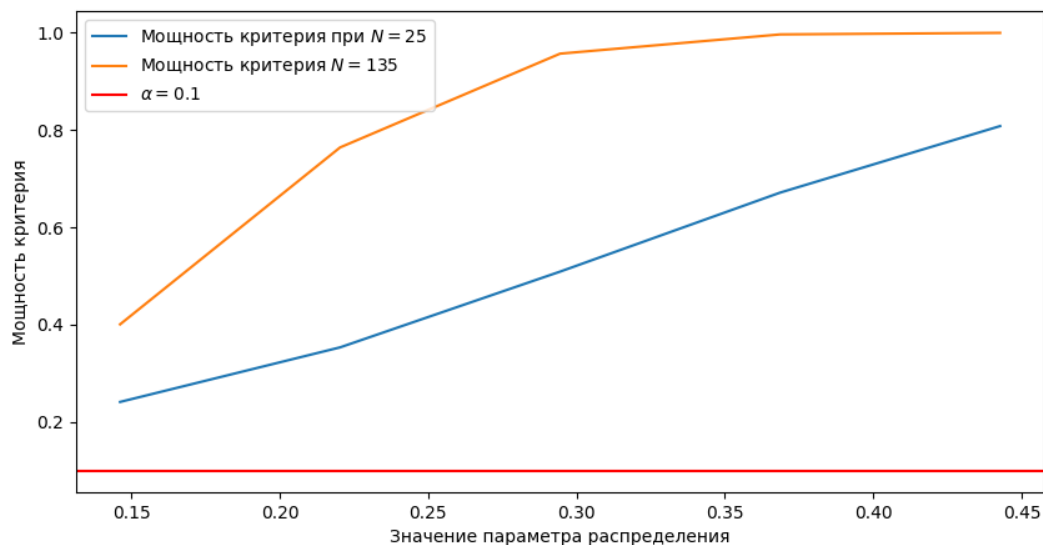


Рисунок 3 – График мощности критерия значимости для малой и полной выборок

Задание 2

Постановка задачи

Используя следующую последовательность шагов, при выполнении проверки гипотезы, решите задачу индивидуального варианта.

1. Опишите характеристику, гипотеза о которой должна быть проверена.
2. Сформулируйте нулевую гипотезу H_0 .
3. Сформулируйте альтернативную гипотезу H_1 .
4. Выберите уровень значимости.
5. Укажите статистику теста, которая будет использоваться, с заменой гипотетического значения, указанного на шаге 2, но без каких-либо вычислений на этом этапе.
6. Убедитесь, что все предположения, необходимые для теста, являются разумными.
7. Вычислите все величины, фигурирующие в статистике теста, а затем значение самой статистики теста.

8. Определите Р-значение, связанное с наблюдаемым значением тестовой статистики. Р-значение (также иногда называемое наблюдаемым уровнем значимости) является мерой несоответствия между предполагаемым значением для характеристики генеральной совокупности и наблюдаемой выборкой. Эта вероятность, при условии, что H_0 истинно, получить значение тестовой статистики, по крайней мере, столь же несовместимое с H_0 , как то, что наблюдалось.

9. Сформулируйте вывод (который состоит в том, чтобы отклонить H_0 , если Р-значение $\leq \alpha$, и не отклонять H_0 в противном случае). Затем следует сформулировать вывод в контексте проблемы и указать уровень значимости.

Вариант 14

Медицинские исследования показали, что многократное разгибание запястья более чем на 20 градусов повышает риск травм запястья и кисти. Каждый из 24 студентов Корнельского университета использовал предложенную новую конструкцию компьютерной мыши, и во время использования мыши фиксировалось разгибание запястья каждого студента. Приведены данные, соответствующие суммарным значениям, приведенным в статье "Сравнительное исследование двух конструкций компьютерных мышей" (Cornell Human Factors Laboratory Technical Report).

27 28 24 26 27 25 25 24 24 24 25 28 22 25 24 28 27 26 31 25 28 27 27 25

Используйте эти данные для проверки гипотезы о том, что среднее значение разгибания запястья у людей, использующих новый дизайн мыши, больше 20 градусов. Уровень значимости 0,05.

Листинг программы

```
from scipy import stats

# Данные
data = [
    27, 28, 24, 26, 27, 25, 25, 24, 24, 24, 25, 28, 22, 25, 24, 28, 27, 26,
    31, 25, 28, 27, 27, 25
]

alpha = 0.05
null_hypothesis_mean = 20
```

```

# Проводим одновыборочный t-тест
t_statistic, p_value = stats.ttest_1samp(
    data, null_hypothesis_mean, alternative='greater'
)

# Печатаем результаты
print('t-статистика:', t_statistic)
print('p-значение:', p_value)

# Проверяем уровень значимости
if p_value < alpha:
    print(
        'Отвергаем нулевую гипотезу: среднее значение разгибания запястья '
        'больше 20 градусов.'
    )
else:
    print(
        'Не отвергаем нулевую гипотезу: среднее значение разгибания запястья '
        'не превышает 20 градусов.'
    )

```

Результат выполнения задания

Исследуется среднее значение разгибания запястья у студентов, использующих новый дизайн мыши.

Формулирование нулевой гипотезы: среднее значение разгибания запястья у студентов, использующих новый дизайн мыши больше 20 градусов. $H_0 - \mu = 20$.

Формулирование альтернативной гипотезы: среднее значение разгибания запястья у студентов, использующих новый дизайн мыши меньше или равно 20 градусам. $H_1 - \mu \geq 20$.

Уровень значимости $\alpha = 0.05$.

Будем использовать одновыборочный -тест. Заменяем гипотетическое значение в тесте.

Используем stats.ttest_1samp для вычисления -статистики и p -значения.

P -значение является вероятностью получить значение -статистики, как минимум такое несовместимое с H_0 , как наблюдаемое, при условии, что H_0 верна.

```
t-статистика: 14.832588870444045
p-значение: 1.4467722904185616e-13
Отвергаем нулевую гипотезу: среднее значение разгибания запястья больше 20 градусов.
```

Рисунок 4 – Результат выполнения программы

Задание 3

Постановка задачи

Пусть проверяется простая гипотеза относительно закона распределения для выборки из ПЗ№2 $H_0: F(x) = F_{\text{теор}}(x)$, при различных уровнях значимости α и для различных объемов выборки.

1. объем $k_1 < 20$ (любые k_1 значений из заданной выборки);
2. объем $k_2 = N$ (полный объем исходной выборки) Используя критерии согласия Колмогорова, ω^2 (Крамера – фон Мизеса), χ^2 Пирсона принять или опровергнуть основную гипотезу.

Листинг программы

```
import argparse
import csv
import random

import numpy as np
from scipy import stats

parser = argparse.ArgumentParser()
parser.add_argument("-file")

args = parser.parse_args()
file = args.file or "../data/Test14.csv"

points = []
with open(file, newline='') as csvfile:
    reader = csv.reader(csvfile, delimiter=' ', quotechar='|')
    for row in reader:
        points.append(float("".join(row)))
points.sort()

def test_hypothesis(data: np.array, alpha: float):
    mean = np.mean(data)
    std_dev = np.std(data)
```



```

theoretical_values = stats.norm(mean, std_dev)

ks_statistic = stats.kstest(
    data,
    theoretical_values.cdf
)

cvm_statistic = stats.cramervonmises(
    data,
    theoretical_values.cdf
)

chi2_statistic = stats.chisquare(
    data
)

print(f'{ks_statistic.statistic=}')
if ks_statistic.statistic < alpha:
    print(
        "Отвергаем основную гипотезу с использованием критерия
Колмогорова-Смирнова"
    )
else:
    print(
        "Принимаем основную гипотезу с использованием критерия
Колмогорова-Смирнова"
    )

print(f'{cvm_statistic.statistic=}')
if cvm_statistic.statistic < alpha:
    print(
        "Отвергаем основную гипотезу с использованием критерия Крамера-
фон-Мизеса"
    )
else:
    print(
        "Принимаем основную гипотезу с использованием критерия Крамера-
фон-Мизеса"
    )

print(f'{chi2_statistic.statistic=}')
if chi2_statistic.statistic < alpha:
    print("Отвергаем основную гипотезу с использованием критерия
Пирсона")
else:
    print("Принимаем основную гипотезу с использованием критерия
Пирсона")

if __name__ == '__main__':
    data = np.array(points)
    sample = random.sample(list(points), 13)
    alpha = 0.001
    print(f"k_1={len(data)}")
    test_hypothesis(data, alpha)
    print(f"k_2={len(sample)}")
    test_hypothesis(sample, alpha)

```

Результат выполнения задания

Используемые критерии значимости: 0.05,0.01,0.001.

Используется полная выборка и 11 случайных значений из выборки.

Для расчета критических значений используется библиотека `scipy.stats`.

Проверим гипотезу о том, что заданная выборка является выборкой из нормального закона распределения.

```
k_1=135
ks_statistic.statistic=0.07791338453392271
Принимаем основную гипотезу с использованием критерия Колмогорова-Смирнова
cvm_statistic.statistic=0.08340505551000434
Принимаем основную гипотезу с использованием критерия Крамера-фон-Мизеса
chi2_statistic.statistic=1390.6559372557554
Принимаем основную гипотезу с использованием критерия Пирсона
k_2=13
ks_statistic.statistic=0.13841247258877248
Принимаем основную гипотезу с использованием критерия Колмогорова-Смирнова
cvm_statistic.statistic=0.0401856228366958
Принимаем основную гипотезу с использованием критерия Крамера-фон-Мизеса
chi2_statistic.statistic=33.920287364010534
Принимаем основную гипотезу с использованием критерия Пирсона
```

Рисунок 5 – Значения статистики для уровня значимости $\alpha = 0.01$

```
k_1=135
ks_statistic.statistic=0.07791338453392271
Принимаем основную гипотезу с использованием критерия Колмогорова-Смирнова
cvm_statistic.statistic=0.08340505551000434
Принимаем основную гипотезу с использованием критерия Крамера-фон-Мизеса
chi2_statistic.statistic=1390.6559372557554
Принимаем основную гипотезу с использованием критерия Пирсона
k_2=13
ks_statistic.statistic=0.22668782774994561
Принимаем основную гипотезу с использованием критерия Колмогорова-Смирнова
cvm_statistic.statistic=0.12173571956981535
Принимаем основную гипотезу с использованием критерия Крамера-фон-Мизеса
chi2_statistic.statistic=184.3192584478578
Принимаем основную гипотезу с использованием критерия Пирсона
```

Рисунок 6 – Значения статистики для уровня значимости $\alpha = 0.05$

```
k_1=135
ks_statistic.statistic=0.07791338453392271
Принимаем основную гипотезу с использованием критерия Колмогорова-Смирнова
cvm_statistic.statistic=0.08340505551000434
Принимаем основную гипотезу с использованием критерия Крамера-фон-Мизеса
chi2_statistic.statistic=1390.6559372557554
Принимаем основную гипотезу с использованием критерия Пирсона
k_2=13
ks_statistic.statistic=0.1448636045529269
Принимаем основную гипотезу с использованием критерия Колмогорова-Смирнова
cvm_statistic.statistic=0.03410654018172439
Принимаем основную гипотезу с использованием критерия Крамера-фон-Мизеса
chi2_statistic.statistic=-35.54760588646404
Отвергаем основную гипотезу с использованием критерия Пирсона
```

Рисунок 7 – Значения статистики для уровня значимости $\alpha = 0.001$

Вывод: в ходе выполнения домашней работы были сформированы практические навыки выбора подходящих методов обработки информации, исходя из требований технического задания; генерации различных вариантов решений поставленных задач и выбора наиболее рациональных вариантов.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Гмурман, В.Е. Теория вероятностей и математическая статистика: учеб. пособие для вузов/ В.Е. Гмурман. - М.: Юрайт, 2014. – 479 с. 21
2. Гринь, А.Г. Вероятность и статистика [Электронный ресурс]: учебное пособие/ А.Г. Гринь.— Омск: Омский государственный университет им. Ф.М. Достоевского, 2013.— 304 с.— Режим доступа: <http://www.iprbookshop.ru/24879.html>
3. Кельберт М.Я. Вероятность и статистика в примерах и задачах [Электронный ресурс]/ Кельберт М.Я. Сухов Ю.М.. - М.: МЦНМО, 2010. - Т. 1. Основные понятия теории вероятностей и математической статистики. - 486 с. - URL: [//biblioclub.ru/index.php?page=book&id=69109](http://biblioclub.ru/index.php?page=book&id=69109)