САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Направление: 01.03.02 Прикладная математика и информатика
ООП: Прикладная математика, фундаментальная информатика и программирование

ОТЧЕТ ПО УЧЕБНОЙ ПРАКТИКЕ

(НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ)

Тема задания: Проверка статистических гипотез на данных ЕМИСС и EUROSTAT

Выполнил: Бударов Алексей Андреевич, студент группы 21Б07

Руководитель: Панкратова Ярославна Борисовна

Санкт-Петербург
2023

СОДЕРЖАНИЕ

ПОСТАНОВКА ЗАДАЧИ

Целью данной научно-исследовательской работы является ознакомление и изучение статистических критериев с целью их применения для проверки статистических гипотез на реальных данных.

Для реализации поставленной задачи необходимо:

1. Провести обзор основных статистических критериев.
2. Изучить принципы работы каждого критерия и его область применения.
3. Собрать или использовать реальные данные, подходящие для статистического анализа.
4. Сформулировать статистические гипотезы, которые могут быть проверены с использованием выбранных критериев.
5. Применить соответствующие статистические критерии для проверки гипотез.
6. Интерпретировать результаты тестов и сделать выводы относительно подтверждения или отвержения гипотез.

Понимание и умение применять статистические критерии является ключевым навыком в анализе данных. Это позволяет исследователям делать выводы на основе данных, а также проверять статистическую значимость различий и зависимостей, что важно в научных и практических исследованиях.

ТЕОРЕТИКО-МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ ИССЛЕДОВАНИЯ

Статистическая гипотеза – любое предположение о законе распределения генеральной совокупности.
Выдвинутую изначально гипотезу, подлежащую статистической проверке, называют нулевой и обозначают H0
Гипотезу, альтернативную нулевой, называют альтернативной и обозначают H1 [1]

Статистический критерий – это метод статистического сопоставления высказанной нулевой гипотезы с имеющимися выборочными данными, сопровождаемый количественной оценкой достоверности получаемого вывода. [1]

Статистика критерия – измеримая функция от выборочных данных, на основании численного значения которой принимается решение об отклонении или принятии нулевой гипотезы. [1]

Ошибка первого рода - это ситуация, когда H0 отвергается, хотя она, на самом деле, верна
Ошибка второго рода - это ситуация, когда H0 принимается, хотя она неверна.

Буквой α обозначается уровень значимости или вероятность ошибки первого рода. Буквой β - вероятность ошибки второго рода.

Критической областью называется область значений статистики критерия, при которых отвергается H0. Критические значения - это граница критической области
Минимальный уровень значимости (p _value) - это минимальное значение α, при котором основная гипотеза ещё отвергается. [2]

В зависимости от проверяемой нулевой гипотезы выбираются подходящие статистические критерии. В данной научно-исследовательской работе были рассмотрены и использованы такие критерии как:

- Критерий согласия: Колмогорова
- Критерии однородности: Ранговый критерий однородности Вилкоксона, Манна-Уитни
- Критерии равенства средних и дисперсий: Критерий Стьюдента, Фишера-Снедекора
- Критерии о значении параметров распределения: Критерий знаков

Для применения критериев равенства средних и дисперсий при проверке статистических гипотез необходимо, чтобы выборки принадлежали нормально распределенным генеральным совокупностям, что не всегда выполняется на реальных статистических данных. Поэтому в данной научно-исследовательской работе были рассмотрены два датасета (из ЕМИСС и EUROSTAT) для полного охвата статистических критериев

АНАЛИЗ ДАННЫХ, ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

В ходе работы были использованы библиотеки Python такие как:

- Matplotlib, Seaborn – для визуализации данных
- Pandas, NumPy - для начальной предобработки данных
- Scipy, Statsmodels – библиотеки, включающие в себя основные статистические критерии и прочие инструменты для непосредственной проверки статистических гипотез.

ДАННЫЕ ЕМИСС – ПРЕДОБРАБОТКА И АНАЛИЗ

Первый набор данных был выбран из единой межведомственной информационной-статистической системы (ЕМИСС).
Был выбран показатель - Среднее отработанное время в неделю занятыми на основной работе по полу и возрасту.

Структура данных:
Для расширения выборки были объединены по годам (от 2019 до 2022 года) показатели среднего времени работы в неделю у мужчин и женщин разных возрастов. В данной работе уделялось внимание именно разделению по половому признаку, поэтому были объединены также и возрастные показатели.

После предобработки исходных данных получили датасет состоящий из:

1) Среднего времени работы мужчин
2) Среднего времени работы женщина
3) Среднего времени работы обоих полов

Среднее значение, стандартное отклонение, для каждой из выборок

```
df_both.describe()
```

```
count    88.000000
mean     35.976136
std       2.882886
min      27.400000
25%      34.400000
50%      36.700000
75%      37.900000
max      39.500000
Name: Both, dtype: float64
```

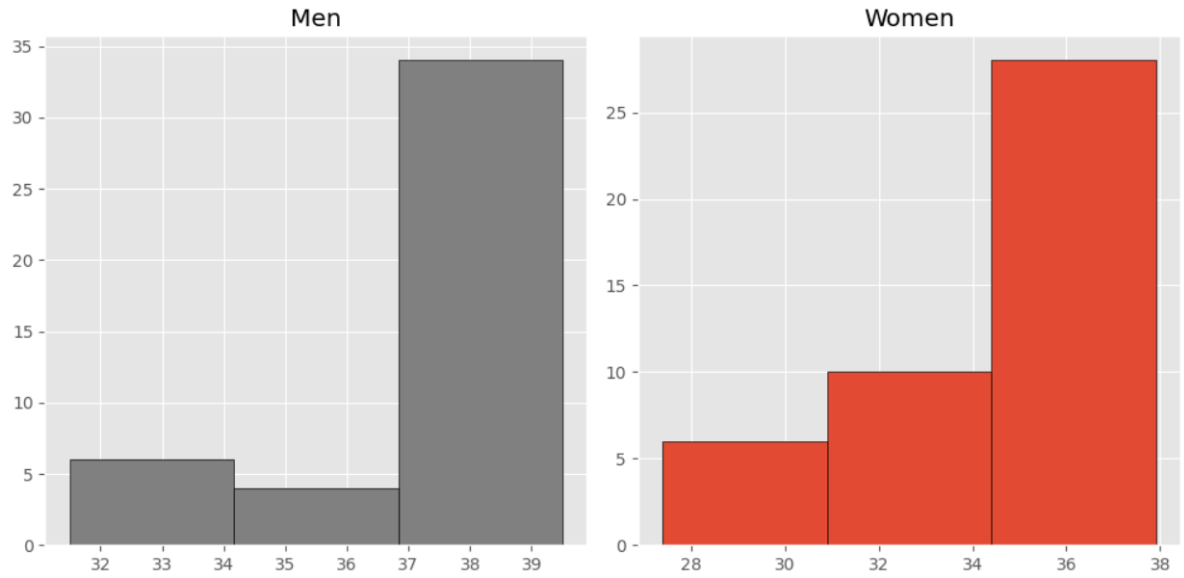
```
df_men.describe()
```

```
count    44.000000
mean     37.429545
std       2.277451
min      31.500000
25%      37.400000
50%      37.900000
75%      39.125000
max      39.500000
Name: Men, dtype: float64
```

```
df_women.describe()
```

```
count    44.000000
mean     34.522727
std       2.702872
min      27.400000
25%      33.200000
50%      35.500000
75%      36.050000
max      37.900000
Name: Women, dtype: float64
```

Построили гистограммы для среднего времени работы мужчин и женщин для визуального определения типа распределения данных.



Из построенной гистограммы ясно, что данные распределены не нормально. Это исключает возможность применять некоторые критерии для проверки статистических гипотез.

Несмотря на это, похоже, данные принадлежат одному и тому же закону распределения. Это мы можем проверить с помощью рангового критерия однородности Вилкоксона, и критерия Манна-Уитни.

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Для ясности и простоты изложения введем обозначения:

α – критерий значимости = 0.05,

F_men – функция распределения среднего времени работы мужчин

F_women - функция распределения среднего времени работы женщин

F_both - функция распределения среднего времени работы обоих полов

H0: F_men = F_women

Критерий Вилкоксона:

```
stats.wilcoxon(np.log(df_men), np.log(df_women), alternative = 'two-sided')
```

```
WilcoxonResult(statistic=0.0, pvalue=1.1368683772161603e-13)
```

P_value < α – отвергаем нулевую гипотезу

Осуществим сдвиг выборки среднего времени работы женщин на разницу между средними значениями рассматриваемых выборок и проверим ту же нулевую гипотезу

```
shift = abs(df_men.describe()['mean'] - df_women.describe()['mean'])

stats.wilcoxon(np.log(df_men), np.log(df_women+shift), alternative = 'two-sided')
```

```
WilcoxonResult(statistic=478.0, pvalue=0.8488191791855115)
```

P_value > α – не отвергаем нулевую гипотезу.

Критерий Манна-Уитни:

```
stats.mannwhitneyu (np.log10(df_men), np.log10(df_women+shift))
```

```
MannwhitneyuResult(statistic=913.0, pvalue=0.6490181793887331)
```

P_value > α – не отвергаем нулевую гипотезу.

Это означает, что на основе предоставленных данных нельзя сделать статистически обоснованный вывод о том, что распределения выборок различны.

ДАННЫЕ EUROSTAT – ПРЕДОБРАБОТКА И АНАЛИЗ

Второй набор данных был взят из EUROSTAT и представляет собой средний доход стран Европы во временном промежутке от 2013 до 2022 года.

Структура данных:

Аналогично предыдущему набору данных были объединены все значения по возрастам и половой принадлежности.

Средний доход в датасете в евро. Для улучшения анализа данных представим значения среднего дохода стран в тысячах евро.

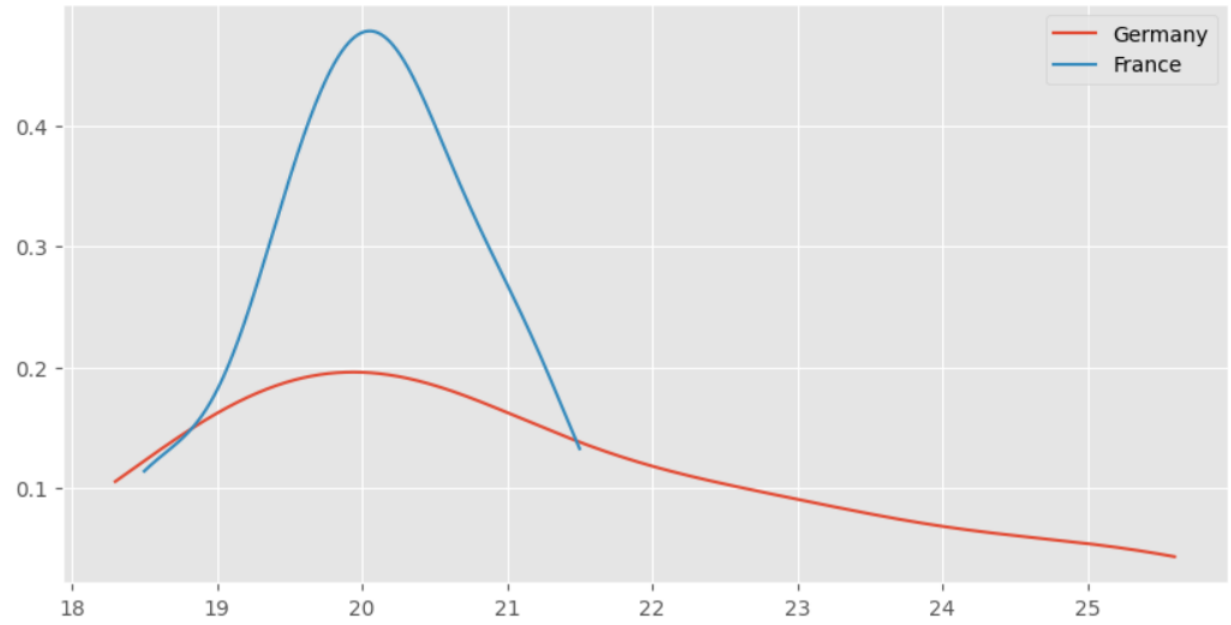
Предобработанный датасет:

Страна – Среднее значение дохода (в тыс. евро) с 2013 до 2022 года

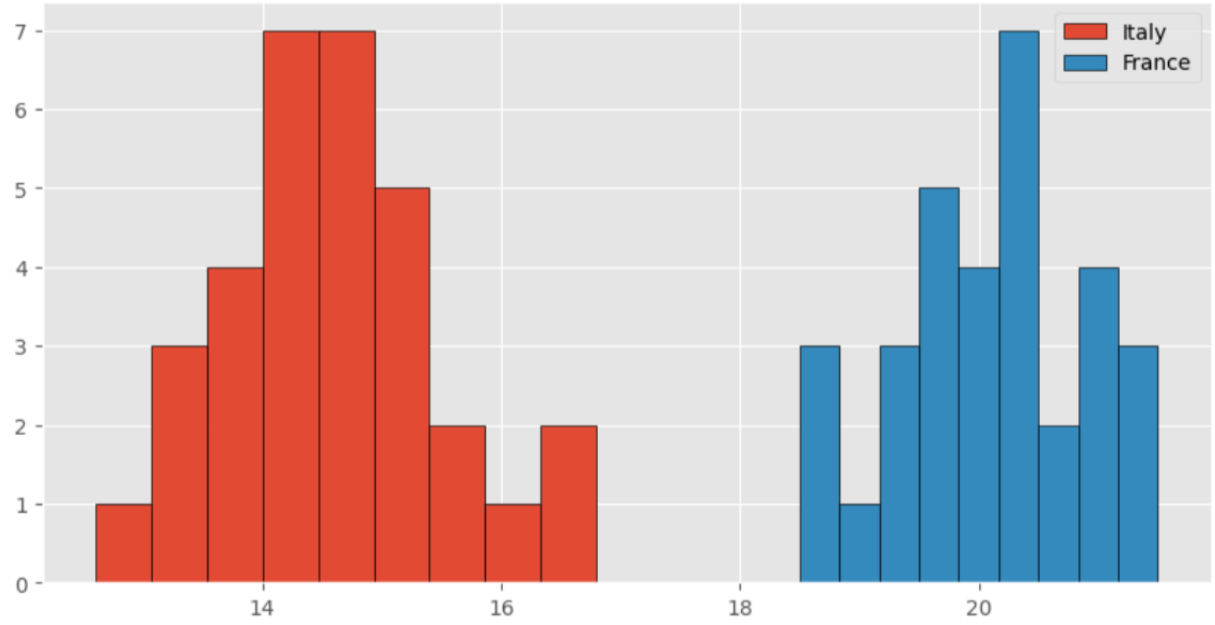
```
Index(['Country', 'Belgium', 'Bulgaria', 'Czechia', 'Denmark', 'Germany',
       'Estonia', 'Ireland', 'Greece', 'Spain', 'France', 'Croatia', 'Italy',
       'Cyprus', 'Latvia', 'Lithuania', 'Luxembourg', 'Hungary', 'Malta',
       'Netherlands', 'Austria', 'Poland', 'Portugal', 'Romania', 'Slovenia',
       'Slovakia', 'Finland', 'Sweden', 'Norway', 'Switzerland', 'Montenegro',
       'North Macedonia', 'Serbia', 'Türkiye'],
      dtype='object')
```

Выберем страны с распределением среднего дохода схожим с нормальным распределением.

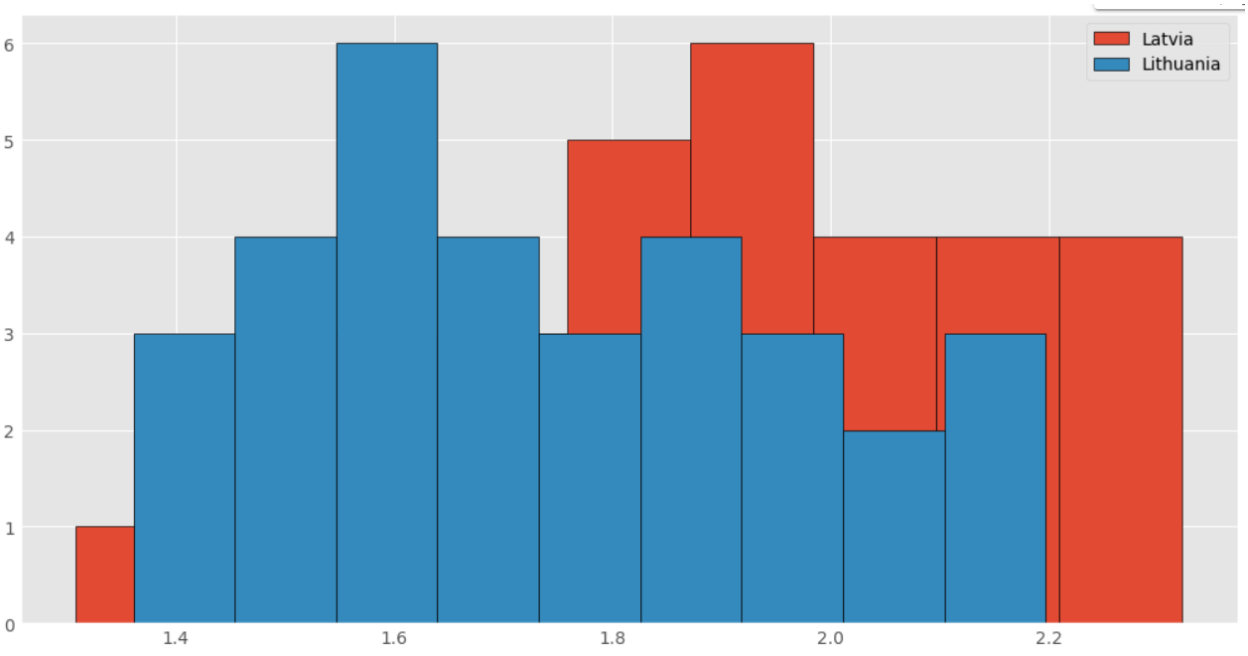
Функции плотности распределения среднего дохода Германии и Франции:



Гистограммы распределения среднего дохода Италии и Франции



Гистограммы распределения среднего дохода Латвии и Литвы



ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Для проверки статистических гипотез выберем несколько стран: Францию, Италию и Германию, Латвию и Литву

Проверим с помощью критерия знаков гипотезу о равенство медианы генеральной совокупности значений среднего дохода жителей Франции некоторому значению tetta = 20 тыс.евро

Критерий знаков:

```
from statsmodels.stats.descriptivestats import sign_test
p_value = sign_test(df_test_2, 20)[1]
print('p_value = ', p_value)
```

```
p_value =  0.8555355519056321
```

P_value > α – не отвергаем нулевую гипотезу.

Для Италии: H0 – медиана генеральной совокупности значений среднего дохода жителей Италии = 14 тыс.евро

```
from statsmodels.stats.descriptivestats import sign_test
p_value = sign_test(df_test_1, 14)[1]
print('p_value = ', p_value)
```

```
p_value =  0.0070003666914999485
```

P_value < α – отвергаем нулевую гипотезу.

Критерий знаков для двух выборок

H0 – медиана генеральной совокупности значений среднего дохода жителей Италии = =медиане генеральной совокупности значений среднего дохода жителей Германии. И пусть критерий значимости alpha = 0.01

```
from statsmodels.stats.descriptivestats import sign_test
p_value = sign_test(df_test_1, df_test_2)[1]
print('p_value = ', p_value)
```

```
p_value =  0.029449373483657837
```

Для α = 0.01

P_value > α – не отвергаем нулевую гипотезу.

Проверим гипотезу о равенстве математических ожиданий среднего дохода жителей Франции и Германии из соответствующих генеральных совокупностей с помощью критерия Стьюдента

Критерий Стьюдента:

```
import numpy as np
from scipy.stats import ttest_ind

res = ttest_ind(df_test_1, df_test_2)
print('p_value = ', res[1] )
```

```
p_value =  7.696868429136046e-33
```

P_value < α – отвергаем нулевую гипотезу.

Проверим гипотезу о равенстве дисперсий двух генеральных совокупностей – Среднего дохода жителей Латвии и Литвы с критерием значимости α = 0.01

Критерий Фишера-Снедекора:

```
from scipy.stats import f_oneway

res = f_oneway(df_test_1, df_test_2)
print('p_value = ', res[1] )
```

```
p_value =  0.021420491841551263
```

P_value > α – отвергаем нулевую гипотезу.

ЗАКЛЮЧЕНИЕ

В ходе изучения и применения статистических критериев для проверки гипотез на реальных данных были достигнуты следующие результаты:

1. Обзор статистических критериев: Проведен обзор основных статистических критериев, что позволило лучше понять их принципы работы и область применения.
2. Изучение и применение критериев: Изучены особенности каждого критерия и выбраны соответствующие критерии для анализа реальных данных.
3. Формулировка и проверка гипотез: Сформулированы статистические гипотезы, направленные на выявление различий и зависимостей в реальных данных. Гипотезы успешно проверены с использованием выбранных статистических критериев.
4. Выводы и интерпретация результатов: Полученные результаты анализа данных позволяют сделать выводы о статистической значимости и подтверждении или отвержении поставленных гипотез. Эти выводы имеют важное значение для понимания характеристик и закономерностей данных.

В целом, выполненная работа позволяет лучше освоить инструментарий статистического анализа, что является ключевым элементом в исследованиях и принятии обоснованных решений на основе данных. Освоенные методы могут быть применены в различных областях, где требуется статистическое обоснование выводов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Буре В. М., Парилина Е. М. Теория вероятностей и математическая статистика. СПб.: Лань, 2023

[2] Курбацкий А. Н. Лекции по теории вероятностей. Статистическая гипотеза. Проверка гипотез. МШЭ МГУ 2020 [Электронный ресурс] : URL: https://mse.msu.ru/wp-content/uploads/2020/03/Лекция-6-Проверка-гипотез.pdf.

[3] Ю.Н.Тюрин, А.А.Макаров АНАЛИЗ ДАННЫХ НА КОМПЬЮТЕРЕ. Москва, Издательство МЦНМО,2016

[4] Computer Science Center : Лекции по анализу данных. [Электронный ресурс]: URL: https://youtube.com/playlist?list=PLlb7e2G7aSpRb95_Wi7lZ-zA6fOjV3_l7&si=_P1as40xkkHJySW4

[5] Python и статистический вывод: часть 3. [Электронный ресурс]: URL: https://habr.com/ru/articles/556852/

[6] Understanding Statistical Hypothesis Testing: The Logic of Statistical Inference [Электронный ресурс]: URL: https://www.mdpi.com/2504-4990/1/3/54#

[7] Буре В. М., Парилина Е. М., Седаков А.А. Методы прикладной статистики в R и Excel СПб.: Лань, 2016

[8] Выбор статистического критерия для проверки гипотез А.М.Гржибовский 2008 [Электронный ресурс]:URL: https://cyberleninka.ru/article/n/vybor-statisticheskogo-kriteriya-dlya-proverki-gipotez/viewer

[9] SciPy Statistical Function documentation [Электронный ресурс]: URL:https://docs.scipy.org/doc/scipy/reference/stats.html

[10] David J. Sheskin Handbook of Parametric and Nonparametric Statistical Procedures 2000 by Chapman & Hall/CRC