МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Практикум № 4

з курсу «Аналіз даних в інформаційних системах»

на тему: «Вивідна статистика»

|  |  |
| --- | --- |
| Викладач:  Ліхоузова Т. А. | Виконав:  студент 2 курсу  групи ІП-11 ФІОТ  Головня О. Р. |

Київ-2023

**Основне завдання** Приклад виконання

Скачати дані файлу Data2.csv.

1. Подивитись, проаналізувати структуру
2. Вказати, чи є параметри, що розподілені за нормальним законом
3. Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів
4. Вказати, в якому регіоні розподіл викидів СО2 найбільш близький до нормального
5. Побудувати кругову діаграму населення по регіонам

Для виконання лабораторної роботи було використано Python.

(За основу взято третю лабу)

Код програми:

import pandas as pd

import matplotlib.pyplot as plt

from scipy.stats import shapiro

from scipy.stats import ttest\_1samp

import seaborn as sns

# зчитування файлу та запис даних у DataFrame

df = pd.read\_csv("Data2.csv", delimiter=";")

# Зміна типів даних у стовпцях

df['Country Name'] = df['Country Name'].astype('string')

df['Region'] = df['Region'].astype('string')

df['GDP per capita'] = df['GDP per capita'].str.replace(',', '.').astype(float)

df['GDP per capita'] = df['GDP per capita'].astype('float')

mean\_populatiion = df['Populatiion'].mean()

df['Populatiion'] = df['Populatiion'].fillna(mean\_populatiion).astype(int)

df['Populatiion'] = df['Populatiion'].astype('int')

df['CO2 emission'] = df['CO2 emission'].str.replace(',', '.').astype(float)

df['CO2 emission'] = df['CO2 emission'].astype('float')

df['Area'] = df['Area'].str.replace(',', '').astype(int)

df['Area'] = df['Area'].astype('int')

# замінюємо від'ємні значення в стовпці Populatiion на середнє значення

mean\_populatiion = df['Populatiion'].mean()

df.loc[df['Populatiion'] < 0, 'Populatiion'] = mean\_populatiion

# замінюємо від'ємні значення в стовпці Сфзшеф на середнє значення

mean\_GDP = df['GDP per capita'].mean()

df.loc[df['GDP per capita'] < 0, 'GDP per capita'] = mean\_GDP

# замінюємо від'ємні значення в стовпці CO2 emission на середнє значення

mean\_CO2 = df['CO2 emission'].mean()

df.loc[df['CO2 emission'] < 0, 'CO2 emission'] = mean\_CO2

# замінюємо від'ємні значення в стовпці Area на середнє значення

mean\_area = df['Area'].mean()

df.loc[df['Area'] < 0, 'Area'] = mean\_area

# замінюємо NaN значення середнім значенням

mean\_GDP = df['GDP per capita'].mean()

df['GDP per capita'] = df['GDP per capita'].fillna(mean\_GDP)

# замінюємо NaN значення середнім значенням

mean\_CO2 = df['CO2 emission'].mean()

df['CO2 emission'] = df['CO2 emission'].fillna(mean\_CO2)

# заміна NaN значень у стовпці "Area" на середнє значення

mean\_area = df['Area'].mean()

df['Area'] = df['Area'].fillna(mean\_area).astype(int)

# GDP per capita

stat, p = shapiro(df['GDP per capita'])

alpha = 0.05

if p > alpha:

print("GDP per capita is normally distributed")

else:

print("GDP per capita is not normally distributed")

# Population

stat, p = shapiro(df['Populatiion'])

alpha = 0.05

if p > alpha:

print("Population is normally distributed")

else:

print("Population is not normally distributed")

# CO2 emission

stat, p = shapiro(df['CO2 emission'])

alpha = 0.05

if p > alpha:

print("CO2 emission is normally distributed")

else:

print("CO2 emission is not normally distributed")

# Area

stat, p = shapiro(df['Area'])

alpha = 0.05

if p > alpha:

print("Area is normally distributed")

else:

print("Area is not normally distributed")

# вибірка

sample = df['GDP per capita']

# перевірка гіпотези про рівність середнього і медіани

t\_stat, p\_val = ttest\_1samp(sample, sample.median())

# виведення результату

if p\_val < 0.05:

print("Різниця між середнім та медіаною значима")

else:

print("Різниця між середнім та медіаною не є значимою")

# Розділення даних на регіони

regions = df['Region'].unique()

# Для кожного регіону:

for region in regions:

# Відбір даних для поточного регіону

data = df[df['Region'] == region]['CO2 emission']

# Побудова гістограми

data.plot.hist(alpha=0.5, bins=30)

# Перевірка нормальності розподілу

stat, p = shapiro(data)

if p > 0.05:

print(f"{region}: розподіл є нормальним (p-value = {p:.3f})")

else:

print(f"{region}: розподіл не є нормальним (p-value = {p:.3f})")

# Обчислення сумарного населення за регіонами

pop\_by\_region = df.groupby("Region")["Populatiion"].sum()

# Побудова кругової діаграми

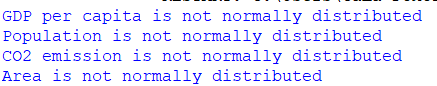
plt.pie(pop\_by\_region, labels=pop\_by\_region.index, autopct='%1.1f%%')

plt.title("Населення країн за регіонами")

plt.show()

1. **Основне завдання:**

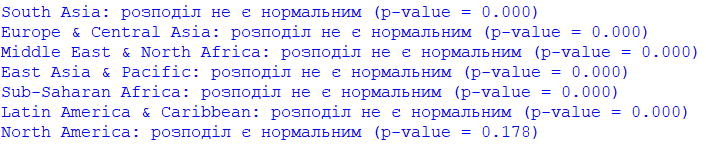
Вказати, чи є параметри, що розподілені за нормальним законом:



Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів:



Вказати, в якому регіоні розподіл викидів СО2 найбільш близький до нормального:



Побудувати кругову діаграму населення по регіонам:

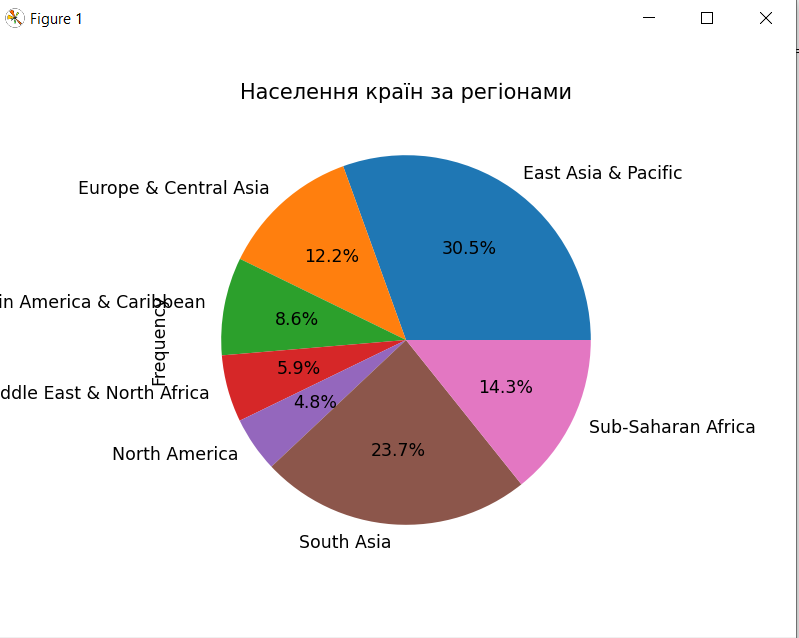


Рисунок 1.1. Kруговa діаграмa населення по регіонам.

1. **Додаткове завдання:**

**Висновок:**

Під час виконання лабораторної роботи я проаналізував структуру вказав, чи є параметри, що розподілені за нормальним законом, перевірив гіпотезу про рівність середнього і медіани для одного з параметрів вказав, в якому регіоні розподіл викидів СО2 найбільш близький до нормального та побудував кругову діаграму населення по регіонам.