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-14 ФІОТ
Шляхтун Денис

Київ-2023

Тема: Вивідна статистика.

Мета роботи: ознайомитись з

- методами визначення точкових оцінок параметрів розподілу; дослідити, що впливає на якість точкових оцінок;
- методикою визначення інтервальних оцінок параметрів розподілу; дослідити, що впливає на якість інтервальних оцінок;
- методами перевірки статистичних гіпотез про вигляд закону розподілу; дослідити, що впливає на ширину критичної області.

Основне завдання

Скачати дані із 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 CO2 найбільш близький до нормального
5. Побудувати кругову діаграму населення по регіонам

Додаткове завдання

Завдання 1

1. Завантажити карту України Ukraine.jpg
2. Розмістити бульбашки, що відповідають їх населенню, на довільних 5 містах (статистику взяти в інтернеті)
3. Знайти найбільшу відстань між містами в пікселях та кілометрах

Завдання 3

1. Завантажити shape-файл с областями України.
2. Побудувати картограми для прибутку населення на 1 особу і ВВП по регіонам за 2016 рік.
3. По даним за 2006-2015 роки для кожного регіону розрахувати коефіцієнт кореляції між прибутком населення на 1 особу та ВВП. Відобразити на картограмі.

Виконання основного завдання.

Виконання комп’ютерного практикуму здійснювалося засобами R та RStudio.

1. Подивитись, проаналізувати структуру

Дані, що використовуються у цьому практикумі, також використовувалися в минулій роботі, де виправлялися помилки у них, тому застосуємо виправлення (взято значення по модулю і пропущені було замінено середніми).

```
> str(data)
'data.frame':   217 obs. of  6 variables:
 $ Country.Name  : chr  "Afghanistan" "Albania" "Algeria" "American Samoa" ...
 $ Region        : chr  "South Asia" "Europe & Central Asia" "Middle East & North Africa" "East Asia & Pacific" ...
 $ GDP.per.capita: num  562 4125 3917 11835 36989 ...
 $ Population    : num  34656032 2876101 40606052 55599 77281 ...
 $ CO2.emission  : num  9809 5717 145400 165114 462 ...
 $ Area          : num  652860 28750 2381740 200 470 ...
> summary(data)
 Country.Name          Region          GDP.per.capita       Population         CO2.emission           Area
 Length:217         Length:217         Min.    :   285.7   Min.    :1.110e+04   Min.    :      11   Min.    :       2
 Class :character   Class :character   1st Qu.:  2361.2   1st Qu.:7.956e+05   1st Qu.:    1955   1st Qu.:   10887
 Mode  :character   Mode  :character   Median :  7179.3   Median :6.293e+06   Median :   11562   Median :   93030
                                       Mean   : 13445.6   Mean   :3.432e+07   Mean   :  165114   Mean   :  618844
                                       3rd Qu.: 14428.1   3rd Qu.:2.370e+07   3rd Qu.:   82563   3rd Qu.:  447420
                                       Max.   :100738.7   Max.   :1.379e+09   Max.   :10291927   Max.   :17098250
```

Рис. 1 – аналіз вхідних даних

2. Вказати, чи є параметри, що розподілені за нормальним законом

Для аналізу параметрів на нормальність було застосовано функцію, що використовує критерій Шапіро-Уілка.

```
> shapiro.test(data$GDP.per.capita)

        Shapiro-Wilk normality test

data:  data$GDP.per.capita
W = 0.73067, p-value < 2.2e-16

> shapiro.test(data$Population)

        Shapiro-Wilk normality test

data:  data$Population
W = 0.2171, p-value < 2.2e-16

> shapiro.test(data$CO2.emission)

        Shapiro-Wilk normality test

data:  data$CO2.emission
W = 0.17369, p-value < 2.2e-16

> shapiro.test(data$Area)

        Shapiro-Wilk normality test

data:  data$Area
W = 0.33839, p-value < 2.2e-16
```

Рис. 2 – перевірка параметрів на нормальність

$p < 0.05$ для усіх параметрів, тому жоден з них не розподілений за нормальним законом.

3. Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів

```
> wilcox.test(data$CO2.emission, mu=median(data$CO2.emission), conf.int=T)

        Wilcoxon signed rank test with continuity correction

data:  data$CO2.emission
V = 16221, p-value = 9.766e-07
alternative hypothesis: true location is not equal to 11562.05
95 percent confidence interval:
 23358.79 52439.93
sample estimates:
(pseudo)median
      34076.26

> wilcox.test(data$Area, mu=median(data$Area), conf.int=T)

        Wilcoxon signed rank test with continuity correction

data:  data$Area
V = 15737, p-value = 1.243e-05
alternative hypothesis: true location is not equal to 93030
95 percent confidence interval:
 156350 290195
sample estimates:
(pseudo)median
      216319.7
```

Рис. 3 – перевірка гіпотези для двох параметрів

4. Вказати, в якому регіоні розподіл викидів CO2 найбільш близький до нормального

```
> data2[order(-data2$p),]
# A tibble: 6 x 2
  Region                          p
  <chr>                       <dbl>
1 East Asia & Pacific       7.59e-30
2 Europe & Central Asia     7.59e-30
3 Latin America & Caribbean 7.59e-30
4 Middle East & North Africa 7.59e-30
5 South Asia                7.59e-30
6 Sub-Saharan Africa        7.59e-30
```

Рис. 4 – перевірка розподілу викидів по регіонам на нормальність

5. Побудувати кругову діаграму населення по регіонам

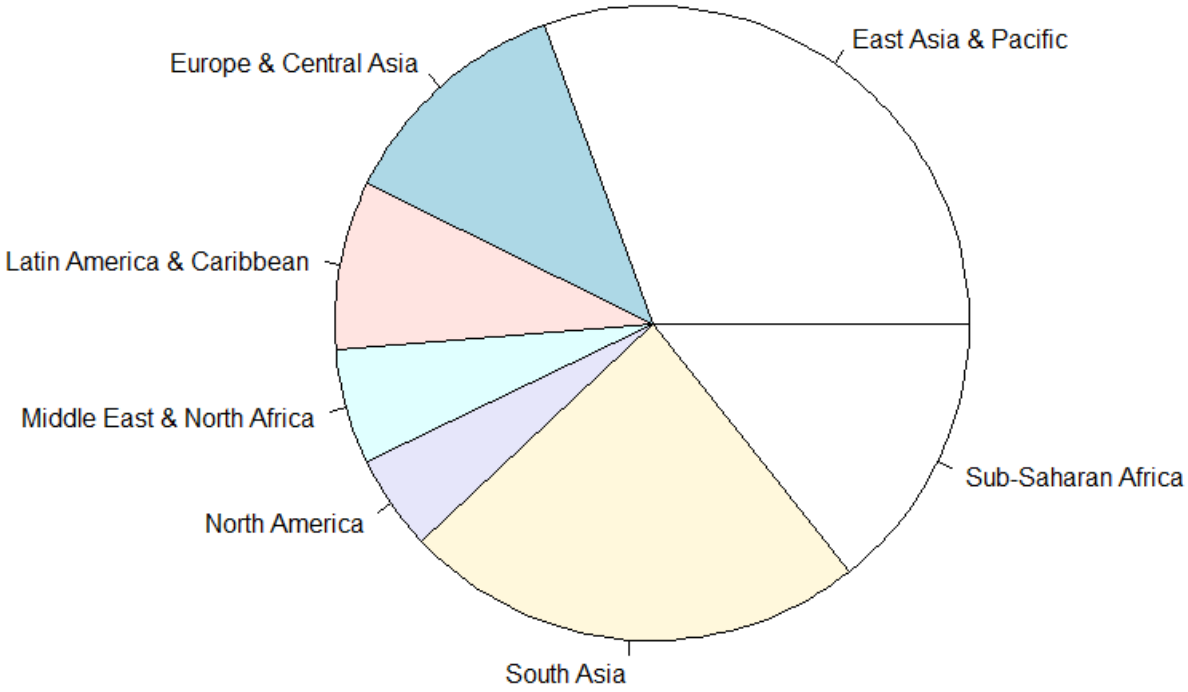

Рис. 5 – кругова діаграму населення по регіонам

Виконання додаткового завдання.

Завдання 1.

Рис. 6 – Карта України з бульбашками, що відображають кількість населення в містах

```
> dist_pixel <- max(dist(data.frame(xy)))
> cat("Найбільша відстань у пікселях:", dist_pixel, "\n")
Найбільша відстань у пікселях: 467.7255
>
> # Відстань між Сумами і Луцьком 660 км
> dist_km <- 660 / dist(data.frame(xy))[1] * dist_pixel
> cat("Найбільша відстань в кілометрах:", dist_km, "\n")
Найбільша відстань в кілометрах: 673.525
```

Рис. 7 – визначення найбільшої відстані в пікселях і кілометрах

Завдання 3.

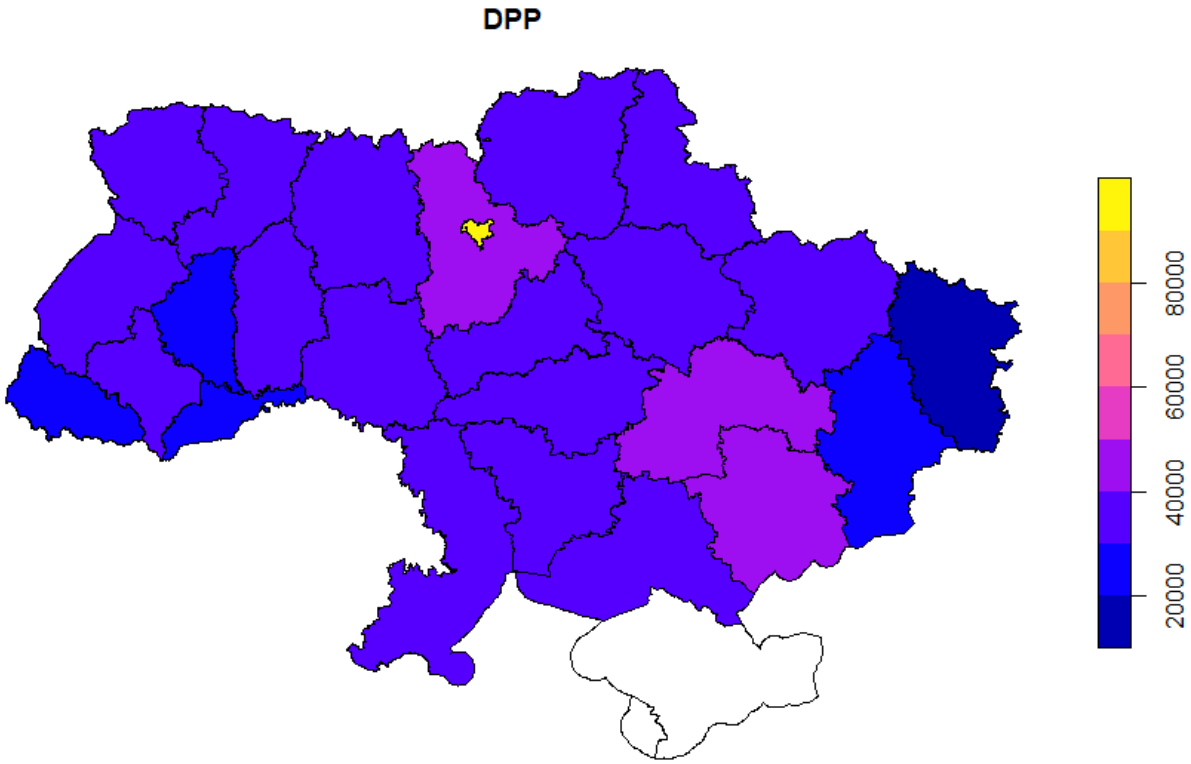

Рис. 7 – картограма прибутку населення на 1 особу по регіонам за 2016 рік

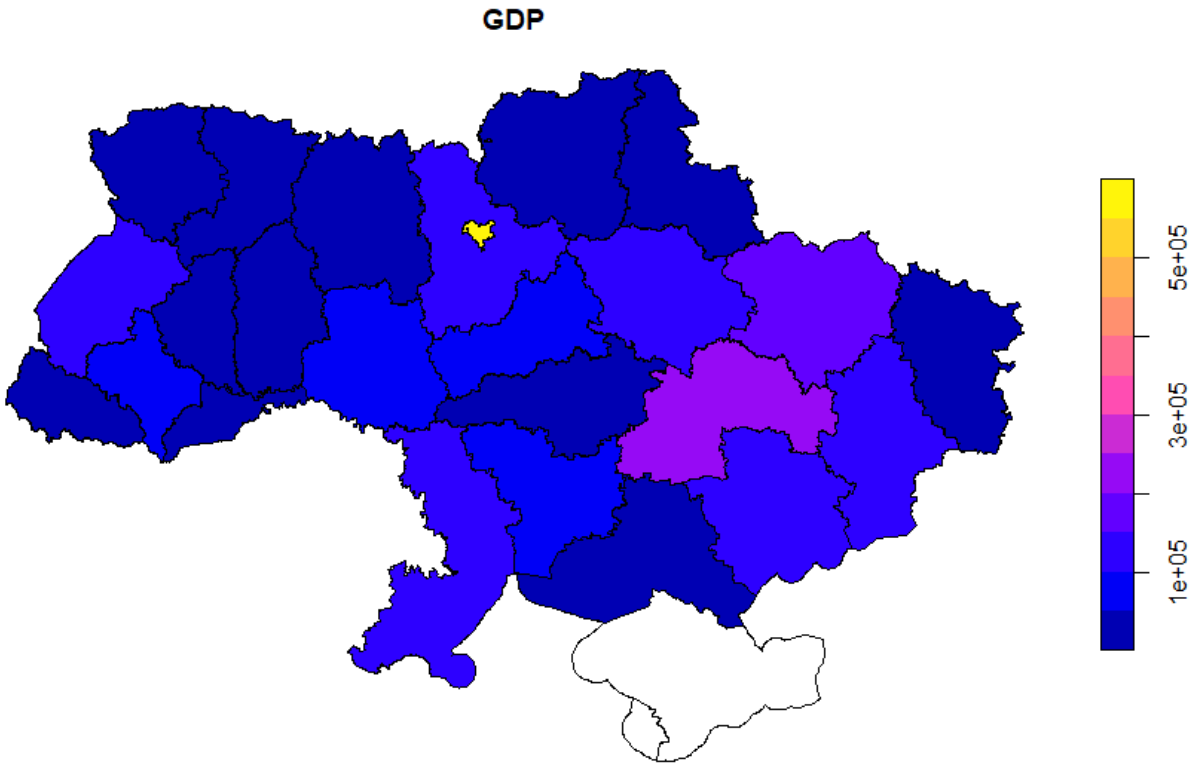

Рис. 8 – картограма Валового регіонального продукту (ВРП) по регіонам за 2016 рік

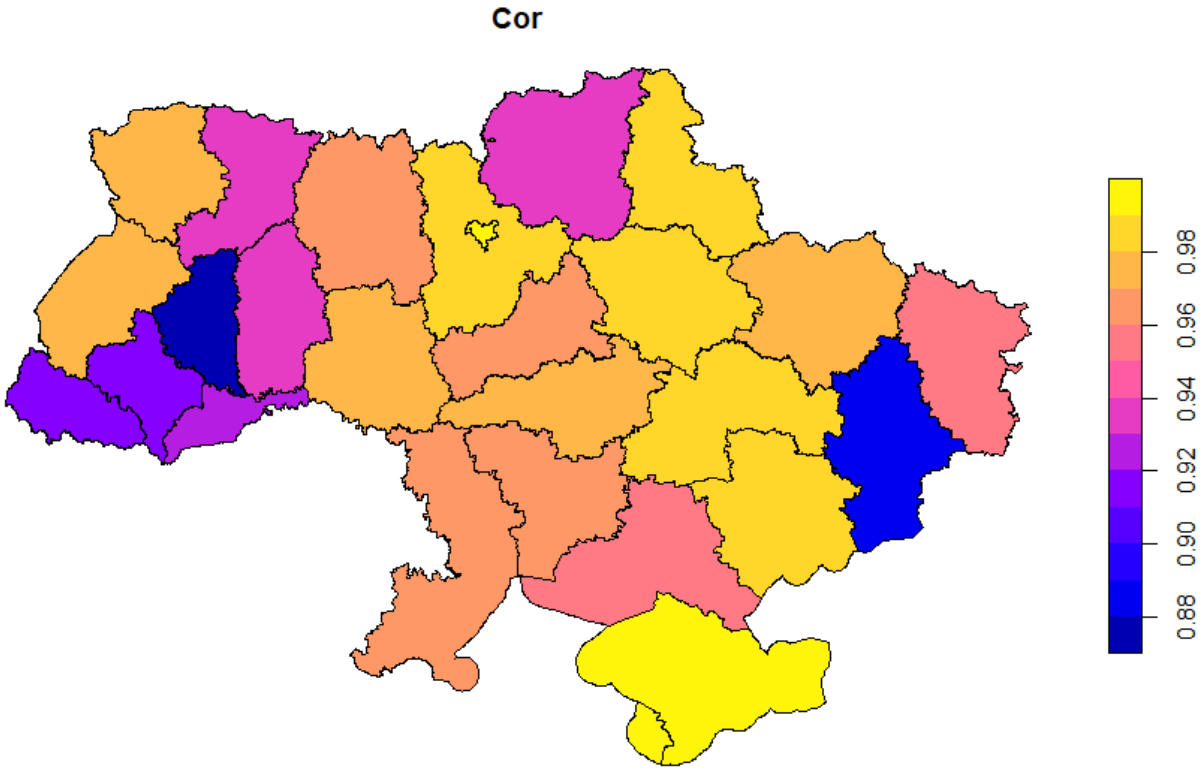

Рис. 9 – коефіцієнт кореляції між прибутком населення на 1 особу та ВВП

Висновок.

При виконанні комп’ютерного практикуму було проаналізовано структуру даних файлу, досліджено параметри, чи розподілені вони за нормальним законом, згруповано дані по регіонам і досліджено окремо викиди кожного регіону на нормальність, було перевірено гіпотезу про рівність середнього і медіани.

Додаток А. Код мовою програмування R

```
# Основне завдання


# 1.Подивитись, проаналізувати структуру
data <- read.csv("Data2.csv", sep=";", header = TRUE, dec = ',')
str(data)

# виправлення помилок в даних

# перейменувати колонку
names(data)[names(data) == "Populatiion"] <- "Population"

# від'ємні значення взяти по модулю
data$GDP.per.capita <- abs(data$GDP.per.capita)
data$Area <- abs(data$Area)

# замінити пропущені значення на середні
data$GDP.per.capita[is.na(data$GDP.per.capita)] <- mean(data$GDP.per.capita, na.rm = TRUE)
data$Population[is.na(data$Population)] <- mean(data$Population, na.rm = TRUE)
data$CO2.emission[is.na(data$CO2.emission)] <- mean(data$CO2.emission, na.rm = TRUE)

str(data)
summary(data)

# 2.Вказати, чи є параметри, що розподілені за нормальним законом
shapiro.test(data$GDP.per.capita)
shapiro.test(data$Population)
shapiro.test(data$CO2.emission)
shapiro.test(data$Area)

# 3.Перевірити гіпотезу про рівність середнього і медіани для одного з параметрів
wilcox.test(data$GDP.per.capita, mu=median(data$GDP.per.capita), conf.int=T)
wilcox.test(data$Population, mu=median(data$Population), conf.int=T)
```

```
wilcox.test(data$CO2.emission, mu=median(data$CO2.emission), conf.int=T)

wilcox.test(data$Area, mu=median(data$Area), conf.int=T)


# 4.Вказати, в якому регіоні розподіл викидів CO2 найбільш близький до нормального

data%>%group_by(Region)%>%summarise(p=shapiro.test(data$CO2.emission)$p.value)

data2 <- data%>%filter(Region != "North
America")%>%group_by(Region)%>%summarise(p=shapiro.test(data$CO2.emission)$p.value)

data2[order(-data2$p),]


# 5.Побудувати кругову діаграму населення по регіонам

diagram <- data%>%group_by(Region)%>%summarise(population=sum(Population))

pie(diagram$population, labels = diagram$Region)



# Додаткове завдання



# Завдання 1


# 1.	Завантажити карту України  Ukraine.jpg

image <- readJPEG("Ukraine.jpg")


par(mar = c(0, 0, 0, 0))

plot(1, xlim = c(0, 831), ylim = c(0, 553), xlab = "", ylab = "")

lim <- par()

rasterImage(image, lim$usr[1], lim$usr[3], lim$usr[2], lim$usr[4])


# 2.	Розмістити бульбашки, що відповідають їх населенню, на довільних 5 містах (статистику взяти в інтернеті)


# Визначити міста і координати

Reg <- c("Суми", "Луцьк", "Вінниця", "Херсон", "Одеса")

xy <- locator(5)


# Розмістити бульбашки, при цьому нормалізувати їхні значення
```

```
Estimates <- c(264753, 217486, 370026,  279131, 1010537)
mycex <- 10 * (Estimates - min(Estimates)) / max(Estimates) + 2
colpts <- rgb(0, 0, 1, 0.7)
points(xy$x, xy$y, cex = mycex,  pch = 21, bg = colpts)

# 3.        Знайти найбільшу відстань між містами в пікселях та кілометрах

dist_pixel <- max(dist(data.frame(xy)))
cat("Найбільша відстань у пікселях:", dist_pixel, "\n")

# Відстань між Сумами і Луцьком 660 км
dist_km <- 660 / dist(data.frame(xy))[1] * dist_pixel
cat("Найбільша відстань в кілометрах:", dist_km, "\n")



# Завдання 3

# 1. Завантажити shape-файл с областями України.

Regions <- read_sf(dsn = "UKR_ADM1.shp")
plot(Regions["Name"])

# 2. Побудувати картограми для прибутку населення на 1 особу і ВВП по регіонам за 2016 рік.

Sys.setlocale("LC_ALL", "C")
GDP<-read.csv("ukr_GDP.csv",sep=';',dec=',', header=T, skip = 1)  # Валовий регіональний продукт
DPP<-read.csv("ukr_DPP.csv",sep=";",dec=",", header=T, skip = 1)  # Прибуток населення на 1 особу

# картограма прибутку населення на 1 особу по регіонам за 2016 рік.
Regions$DPP = NA
for(i in 1:nrow(DPP))
  Regions$DPP[i] <- DPP$X2016[DPP$Name == Regions$Name[i]]
plot(Regions["DPP"])
```

```
# картограма Валового регіонального продукту (ВРП) по регіонам за 2016 рік.
Regions$GDP = NA
for(i in 1:nrow(GDP))
  Regions$GDP[i] <- GDP$X2016[GDP$Name == Regions$Name[i]]
plot(Regions["GDP"])


# 3. По даним за 2006-2015 роки для кожного регіону розрахувати коефіцієнт кореляції між прибутком населення на 1 особу та ВВП. Відобразити на картограмі.
Regions$Cor = NA
for (i in 1:nrow(GDP))
  Regions$Cor[i] <- cor(
          as.numeric(DPP[GDP$Name == Regions$Name[i], 3:12]),
          as.numeric(GDP[GDP$Name == Regions$Name[i], 3:12]),
          use = "complete.obs")
plot(Regions["Cor"])
```