### Zad. 1.

### Otwórz w Excelu plik Roczniaki.xls i przygotuj go w sposób umożliwiający wczytanie zawartych w nim danych do R(tabela z maksymalnie jednym wierszem nagłówka), podając wagę w kilogramach.Zapisz plik w formacie CSV i wczytaj go do R.

zad1 = read.csv("Roczniaki.csv", header = TRUE, sep = ";", check.names = FALSE, dec = ",")

#### a)Wyświetl pierwsze 6 wierszy z odczytanej ramki danych.

zad1[1:6,]

#### b)Wyświetl tylko kolumnę w której jest waga.

zad1[3]

#### c)Wyznacz BMI dla każdego dziecka i zapisz dane pod nazwą bmi.

n = length(zad1$ID)

bmi = c()

for(i in 1:n) {

bmi = c(bmi,c(zad1$`Waga dziecka (w kg)`[i] / ((zad1$`Wzrost dziecka w cm`[i] / 100 \* zad1$`Wzrost dziecka w cm`[i] / 100))))

}

#### d)Posortuj wartości odpowiadające bmi w kolejności nierosnącej.

sort(bmi, decreasing = TRUE)

### Zad. 2.Zainstaluj pakiet foreign i zapoznaj się z funkcją read.spss. Wczytaj za jej pomocą przykładowy plik programu IBM SPSS Statistics o nazwie Employee data.sav. Uzyskaj nazwy zmiennych z tego pliku.

zad2 = read.spss("Employee data.sav", use.value.labels = TRUE, to.data.frame = TRUE, reencode='utf-8')

#### a)Utwórz zmienną poziom\_zarobkow, która przyjmuje następujące wartości:

#### niskie, jeżeli zmienna salary przyjmuje wartości od 0 do 25000 włącznie,

#### przeciętne, jeżeli zmienna salary przyjmuje wartości od 25000 do 35000 włącznie,

#### wysokie, jeżeli zmienna salary przyjmuje wartości powyżej 35000.

n = length(zad2$salary)

poziom\_zarobkow<-array("n", n)

x = as.numeric(levels(zad2$salary))[zad2$salary]

for(i in 1:n) {

ifelse(x[i] <= 25000, poziom\_zarobkow[i]<-"niskie",

ifelse(x[i] > 35000, poziom\_zarobkow[i]<-"wysokie",

poziom\_zarobkow[i]<-"przeciętne"))

}

#### b)Ile jest obserwacji w poszczególnych grupach zdefiniowanych przez zmienną poziom\_zarobków?

dim(poziom\_zarobkow[which(poziom\_zarobkow=="niskie")])

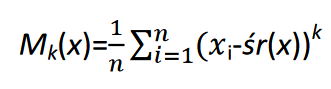
dim(poziom\_zarobkow[which(poziom\_zarobkow=="przeciętne")])

dim(poziom\_zarobkow[which(poziom\_zarobkow=="wysokie")])

#### c)Zapisz zmienną poziom\_zarobków w pliku zarobki.txt.

write.table(poziom\_zarobkow, file = "zarobki.txt", sep=",")

### Zad. 3 Napisz funkcję w R, która biorąc wektor danych liczbowych x oraz liczbę całkowitą k jako argumenty, zwróci k-ty moment centralny wektora x zgodnie ze wzorem:



moment<-function(x, k) {

suma = 0

n = length(x)

for (i in 1:n) {

suma = suma + (x[i]-mean(x))\*\*k

}

return(1 / n \* suma)

}

### Zad. 1. Baza danych LakeHuron zawiera informacje dotyczące corocznego poziomu wody w jeziorze Huron (podane w stopach) w latach 1875-1972.

#### a)Zbuduj szereg przedziałowy dla analizowanych danych:

#### i.o równych długościach przedziałów,

n = length(LakeHuron)

y = cut(LakeHuron, sqrt(n))

#### ii.składający się z 5 klas o(w przybliżeniu)równych liczebnościach.

pkt1 = table(y)

y2 = cut(LakeHuron, breaks = c(575, 577.82, 578.82, 579.37, 579.96, 590))

pkt2 = table(y2)

pkt2

#### b)W jakim przedziale określającym poziom wody w jeziorze znajduje się dominanta?

dominanta = names(sort(pkt1, decreasing = T))[1]

### Zad.2.Baza danych Indometh zawiera informacje na temat stężenia indometacyny (leku przeciwzapalnego) we krwi. Dokonaj analizy zmiennej conc, która zawiera dane liczbowe określające stężenie leku (mcg/ml).

dane = Indometh$conc

#### a)Jaka jest wielkość próbki?

length(dane)

#### b)Z jakiego przedziału przyjmuje wartości analizowana zmienna?

range(dane)

#### c)Skomentuj wzajemne położenie średniej i mediany.

mean(dane)

median(dane)

#Srednia i media znajduja sie w 3 kwartylu.

#### d)Zbadaj symetrię próbki.

library(moments)

skewness(dane)

#Prawostronna asymetria

#### e)Jaka jest średnia obcięta z 50% środkowych obserwacji?

mean(dane, 0.25)

#### f)Zbadaj koncentrację próbki.

kurtosis(dane)

#Dane są skoncentrowane wokół średniej

#### g)Jaka jest długość przedziału, w którym znajduje się 50% środkowych obserwacji?

quantile(dane, c(0.75)) - quantile(dane, c(0.25))

#### h)Ile wynosi poziom stężenia indometacyny, poniżej którego znajduje się 30% obserwacji?

quantile(dane,c(0.3))

#### i)Ile wynosi poziom stężenia indometacyny, powyżej którego znajduje się 20% obserwacji?

quantile(dane,c(0.8))

### Zad. 3 Baza danych iris zawiera dane dotyczące trzech gatunków irysów (gatunek kwiata określa zmienna Species). Porównaj długości kielicha kwiatów irysa (zmienna Sepal.Length) w każdym z tych gatunków wykonując podstawową analizę statystyczną danych.

summary(iris$Petal.Length)

setosa = z[which(iris$Species == "setosa")]

versicolor = z[which(iris$Species == "versicolor")]

virginica = z[which(iris$Species == "virginica")]

summary(setosa)

skewness(setosa)

kurtosis(setosa)

summary(versicolor)

skewness(versicolor)

kurtosis(versicolor)

summary(virginica)

skewness(virginica)

kurtosis(virginica)

LUB

summary(iris$Sepal.Length)

setosa <- iris[1:50,]$Sepal.Length

versicolor <- iris[51:100,]$Sepal.Length

virginica <- iris[101:150,]$Sepal.Length

summary(setosa)

summary(versicolor)

summary(virginica)

### Zad. 1.Załaduj do Rdane zawarte w plikuAnkieta.csv.

dane = read.csv("Ankieta.csv", header=TRUE, sep=";")

#### a.Narysuj wykres pudełkowy dla wieku w rozbiciu na płeć.

attach(dane)

boxplot(wiek~plec)

#### b.Narysuj histogram dla ciśnienia skurczowego.

hist(cisnienie.skurczowe)

#### c.Zmodyfikuj histogram tak, aby na osi pionowej zaznaczone były prawdopodobieństwa empiryczne.

hist(cisnienie.skurczowe, prob=TRUE)

#### d.Odpowiedz na pytanie: czy w ankietowanej grupie było więcej kobiet, czy mężczyzn? rysując odpowiedni wykres.

barplot(table(dane$plec))

#### e.Wyznacz średnie ciśnienie skurczowe dla grup wyznaczonych przez płeć.

e = aggregate(cisnienie.skurczowe, by=list(plec), FUN=mean)

#### f.Zobrazuj otrzymane wyniki na wykresie słupkowym.

barplot(e$x)

### 2.Załaduj do R dane zawarte w pliku UOF\_gs.txt.

plik = read.table("UOF\_gs.txt", header = TRUE)

#### a.Przypisz zmiennym następujące poziomy:

#### i.gender:0 –kobieta,1 –mężczyzna,

plik$gender <- ifelse(plik$gender == 0, "kobieta", "mezczyzna")

#### ii.college: 1 –rolnictwo, 2 –architektura,3 –budownictwo, 4 –administracja,5-leśnictwo, 6–pedagogika,7 –inżynieria,8 –sztuki piękne

plik$college = c("rolnictwo","architektura","budownictwo","administracja", "leśnictwo", "pedagogika", "inżynieria","sztuki piękne")

plik$college <- college[plik$college]

#### Korzystając z pakietu ggplot2 wykonaj następujące wykresy:

#### a.Wykres rozrzutu pensji względem płci

qplot(plik$gender, plik$salary, data=plik)

#### b.Wykres słupkowy/ histogram obrazujący liczebność absolwentów poszczególnych typów uczelni (zmienna college).

g<-ggplot(plik,aes(plik$college))

g+geom\_bar()

#### c.Wykres punktowy zależności rodzaju ukończonych studiów od zarobków, z punktami zróżnicowany mikolorem ze względu na płeć.

g2<-ggplot(plik, aes(plik$college, plik$salary))

g2 + geom\_point(aes(colour = factor(plik$gender)))

#### d.Wykres punktowy zależności rodzaju ukończonych studiów od zarobków, z punktami zróżnicowanymi kształtem ze względu na płeć, a kolorem ze względu na datę ukończenia studiów

g3<-ggplot(plik, aes(plik$college, plik$salary))

g3 + geom\_point(aes(colour = factor(plik$grad\_date)))

### Zad. R.1 Liczbę sprzedanych biletów MZK w Toruniu w kolejnych niedzielach maja i czerwca przedstawia tabelka.



#### Na podstawie tych danych, na poziomie istotności α= 0,1, przetestuj hipotezę,że średnia liczba sprzedawanych biletów w niedziele jest równa 3,2 tys. przeciw hipotezie, że średnia sprzedawanych biletów jest

bilety = c(2900, 3300, 3200, 3200, 3200, 3000, 2900, 3100)

# H0 mu=3200

#### (a) mniejsza niż 3,2 tys.,

#Hipoteza H1: mu1 < 3200

t.test(bilety, mu = 3200, alternative="l")

# Odrzucamy H0, przyjmujemy H1

#### (b) różna od 3,2 tys.,

Hipoteza H1: mu1 != 3200

t.test(bilety, mu = 3200)

# Nie ma podstaw do odrzucanie H0

#### jeżeli wiadomo, że liczba sprzedawanych biletów ma rozkład normalny.

### Zad. R.2 Na pudełkach zapałek jest napisane „średnio 64 zapałki”. W pliku zapalki.txt znajdują się dane dotyczące ilości zapałek w 1000 wylosowanych pudełkach. Na podstawie tych danych zweryfikuj na poziomie istotnościα= 0,05 hipotezę H0:μ= 64 wobec hipotezy alternatywnej H1:μ >64.

# Rozklad nie musi byc normalny

# H0: mu = 64

# H1: mu > 64

t.test(x, mu=64, alternative="g")

# Odrzucamy H0, Przyjmuje H1

### Zad. R.3 Producent płatków mydlanych wysunął hipotezę, że stopień wyprania tkaniny wełnianej płatkami mydlanymi jest wyższy od stopnia wyprania płynem do prania.W celu sprawdzenia tej hipotezy wykonano pomiary stopnia wyprania 10 wycinków tkaniny pranej płatkami, otrzymując w procentach wyniki 74,4; 75,1; 73,0; 72,8;76,2; 74,6; 76,0; 73,4; 72,9; 71,6, oraz 7 wycinków pranych płynem do prania,otrzymując 56,8; 57,8; 54,6; 59,0; 57,1; 58,2; 57,6. Zakładając, że stopień wyprania tkaniny ma rozkład normalny i wiedząc, że test równości wariancji wykonany dla powyższych próbek nie pozwolił na odrzucenie hipotezy zerowej, na poziomieistotnościα= 0,05 zweryfikuj hipotezę wysunietą przez producenta.

#X - procentowy stopien wyprania platkami

#Y - procentowy stopien wyprania plynem

# X i Y niezalezne

# H0: mu = mu2

# H1: mu < mu2

platki = c(74.4, 75.1, 73.0, 72.8, 76.2, 74.6, 76.0, 73.4, 72.9, 71.6)

plyn = c(56.8, 57.8, 54.6, 59.0, 57.1, 58.2, 57.6)

t.test(platki, plyn, alternative="g") **DOPISAĆ** “**var.equal=TRUE”**

# Odrzucamy H0, przyjmujemy H1

# Stopien wyprania platkami jest wyzszy niz plynem

### Zad. R.4 Średnie prędkości tramwaju (w km/h) obliczone dla zmierzonych w środę prędkości 200 tramwajów oraz dla 120 tramwajów w niedzielę znajdują się w plikach tramwajesroda.txt oraz tramwajeniedziela.txt. Na podstawie dostępnych danych zweryfikuj na poziomie istotnościα= 0,05 hipotezę, że średnia prędkość tramwajów w środę jest mniejsza niż w niedzielę.

# H0: mu1 = mu2

# H1: mu1 < mu2

sroda = read.table("tramwaje\_sroda.txt")

niedziela = read.table("tramwaje\_niedziela.txt")

sredniaSroda = sroda$x

sredniaNiedziela = niedziela$x

var.test(sredniaSroda ,sredniaNiedziela )

t.test(sredniaSroda, sredniaNiedziela, alternative="l")

# Odrzucamy H0, przyjmujemy H1

**DOPISAC czy możemy uznać równymi wariancje zmiennych**

### Zad. R.5 Zmierzono ciśnienie tętnicze wśród losowo wybranej grupy chorych na pewną chorobę przed i po podaniu takiego samego leku każdemu z badanych pacjentów.Otrzymano następujące wyniki:

### 

### Na poziomie istotnościα= 0,05 zweryfikuj hipotezę, że stosowany lek nie powoduje zmiany ciśnienia u pacjentów, wobec hipotezy alternatywnej, że wartość przeciętna ciśnienia przed podaniem leku jest wyższa niż po jego podaniu, wiedząc, że ciśnienie tętnicze ma rozkład normalny.

# Zalezne

# H0: mu1 = mu2

# H1: mu1 > mu2

przedLek = c(210, 180, 260, 270, 190, 250, 180, 200)

poLek = c(180, 160, 220, 260, 200, 230, 180, 190)

t.test(przedLek, poLek, paired=TRUE, alternative="g")

# Odrzucamy H0, przyjmujemy H1

**DOPISAC czy możemy uznać równymi wariancje zmiennych**

### Zad. R.6 W czasie sondażu przeprowadzonego przez pracownię badania opinii społecznej spośród 1100 ankietowanych dorosłych Polaków 1090 odpowiedziało, że w ubiegłym miesiącu nie przeczytali żadnej książki, a pozostali potwierdzili, że przeczytaliprzynajmniej jedna książkę. Dane zawierające odpowiedzi na postawione pytanie znajdują się w pliku sondaz.txt. Na ich podstawie, na poziomie istotności 0,01,przetestuj hipotezę, że odsetek dorosłych Polaków, którzy nie przeczytali w ubiegłym miesiącu żadnej książki wynosi 99%, przeciw hipotezie, że odsetek ten jest inny.

#H0: mu = 0.99

#H1: mu != 0.99

dane = read.table("sondaz.txt")

dane = dane$x

doTestu = c()

for(i in 1:1100)

{

if(dane[i] == "nie")

doTestu[i] = 1

else

doTestu[i] = 0

}

t.test(doTestu, mu=0.99)

# Nie ma podstaw do odrzucanie H0

### Zad. R.1 Liczba ocen niedostatecznych uzyskanych na egzaminie z pewnego przedmiotu przez jednakowo liczne grupy studenckie I roku Wydziału Matematyki i Informatyki pewnego uniwersytetu były następujące

### 

### Na poziomie istotności 0,05 testem χ2 zweryfikować hipotezę, że prawdopodobieństwa występowania ocen niedostatecznych w tych grupach są jednakowe.

oceny = c(14, 18, 28, 12, 4, 22, 14, 16, 10, 8, 18, 6, 12)

# H0 - prawdopodobienstwa niedostatecznych podobne

# H1 - prawdopodobienstwa sa rozne

chisq.test(oceny)

# Odrzucamy H0 (p-wartosc jest bardzo mala) rozklad nie jest rownomierny(jednostajny)

# przyjmujemy H1, rozklad nie jest rownomierny

### Zad. R.2 Wyznaczono liczby błędów przy korekcie 500 stronicowej książki. Wyniki opi-sujące liczbę błędów na kolejnych stronach znajdują się w pliku błędy.txt. Na po-ziomie istotności 0,05 zweryfikuj hipotezę, że liczba błędów na stronicy ma rozkładPoissona.

bledy = read.table("bledy.txt")

# H0 - liczba bledow na stronicy ma rozklad Poissona

# H1 - liczba bledow na stronicy nie ma rozkladu Poissona

lambda = mean(bledy$x)

# wartosci {0,1,2,3,4,5,6,7,10}

for (i in 0:7) {

prob[i+1] = dpois(i, lambda)

}

prob[9] = dpois(10,lambda)

chisq.test(table(bledy$x), p = prob, rescale.p = T)

# Nie jestesmy w stanie stwierdzic

# Mala p value, aproksymacja chi-kwadrat moze nie byc niepoprawna, nie jestesmy w stanie stwierdzi poprawnosc H0

### Zad. R.3 Generator liczb losowych wygenerował 30 liczb z rozkładu wykładniczego E(2).Liczby są uporządkowane niemalejąco:

### 

### Za pomocą testu Kołmogorowa na poziomie istotności 0,05 przetestuj zgodność tych danych z rozkładem E(2).

#H0 - dane sa zgodne z rozkladem E(2)

#H1 - dane nie sa zgodne z rozkladem E(2)

dane = c(0.02, 0.03, 0.03, 0.04, 0.04, 0.05, 0.06, 0.11, 0.11,0.16, 0.18, 0.22, 0.24, 0.26, 0.27, 0.36, 0.44, 0.46, 0.46,0.60, 0.65, 0.65, 0.70, 0.80, 0.85, 0.90, 0.95, 1.20, 1.50, 2.00)

ks.test(dane, "pexp", 2)

# p-value = 0.7785 jest duze co swiadczy pozytywnie za hipoteza o zgodnosci z rozkladem wykladniczym E(2)

### Zad. R.4 Wykonano 15 pomiarów czasu likwidowania zrywów na przędzarce obrączko-wej, otrzymując (w s):

### 

### Wykonując odpowiedni test zweryfikuj na poziomie istotności 0,05 hipotezę, że czas likwidacji zrywu ma rozkład N(6,3; (1,5)2).

#H0 Czas likwidacji zrywu ma rozklad N(6,3; (1,5)^2)

#H1 Czas likwidacji zrywu nie ma rozkladu N(6,3; (1,5)^2)

dane = c(4.5, 3.6, 6.0, 6.4, 7.9, 6.9, 6.1, 7.4, 9.0, 4.3, 6.1, 8.2, 4.9, 7.5, 5.8)

shapiro.test(dane)

**lepiej było użyć testu Kolmogorowa-Smirnowa.**

# p-value = 0.9572 jest duze co świadczy pozytywnie o rozkładzie normalnym

### Zad. R.5 Wczytaj do R plik zmienne.txt.

dane=read.table("zmienne.txt")

#### a) Sprawdź, czy którekolwiek ze zmiennych zmienna1, zmienna2, zmienna3 mają te same rozkłady. Posłuż się odpowiednim testem oraz wykresami dystrybuant empirycznych dla wszystkich trzech próbek.

ks.test(dane$zmienna1,dane$zmienna2)

ks.test(dane$zmienna1,dane$zmienna3)

ks.test(dane$zmienna2,dane$zmienna3)

# Dla kazdego testu p-value < 2.2e-16 co świadczy że nie mają tych samych rozkładów

# Wykres dla dystrybuant

plot.ecdf(dane$zmienna1, col="red")

plot.ecdf(dane$zmienna2, col="blue",add=T)

plot.ecdf(dane$zmienna3, col="green",add=T)

#### b) Wykonując odpowiednie testy, spróbuj wyznaczyć konkretne rozkłady, z których pochodzą analizowane próbki.

# Zmienna1

shapiro.test(dane$zmienna1)

# p-value = 0.6906 jest duze co świadczy pozytywnie o rozkładzie normalnym

ks.test(dane$zmienna1, "punif", min(dane$zmienna1), max(dane$zmienna1))

# p-value = 0.001386 jest male co świadczy że mozemy odrzucic hipoteze o rozkladzie jednostkowym

# Zmienna2

shapiro.test(dane$zmienna2)

# p-value = 3.373e-11 jest male co świadczy że mozemy odrzucic hipoteze o rozkladzie normalnym

ks.test(dane$zmienna2, "punif", min(dane$zmienna2), max(dane$zmienna2))

# p-value < 2.2e-16 jest male co świadczy że mozemy odrzucic hipoteze o rozkladzie jednostkowym

ks.test(dane$zmienna2, "pexp", 4)

# p-value = 0.642 jest duze co świadczy pozytywnie o rozkładzie wykladniczym E(4)

# Zmienna3

shapiro.test(dane$zmienna2)

# p-value = 0.0002873 jest male co świadczy że mozemy odrzucic hipoteze o rozkladzie normalnym

ks.test(dane$zmienna3, "punif", min(dane$zmienna3), max(dane$zmienna3))

# p-value = 0.5608 jest duze co świadczy pozytywnie o rozkładzie jednostkowym

# Zatem :

# \* zmienna1 - rozkład normalny

# \* zmienna2 - rozklad wykładniczy E(4)

# \* zmienna3 - rozklad jednostajny

### Kolokwium

dane = read.table("dochody.txt", header = TRUE)

reszta = dane$dochod - 52\*dane$wydatki

summary(reszta)

quantile(reszta,0.6)

library(moments)

skewness(reszta)

kurtosis(reszta)

boxplot(reszta~dane$region)

#H0 ilosć osob zamieszkujacych kazdy region jest tanka sama

t.test(table(dane$region), mu = mean(table(dane$region)) )

# p > 0.05 wiec nie mamy podstaw by odzucic hipoteze H0,(Przyjmujemy H0)

#H0 rozkad zmiennej dochod pochodzi z rozkadu normalnego

shapiro.test(dane$dochod)

#p-value = 0.3302 > 0.05 a wiec nie mamy podstaw by odzucic Ho (Przyjmujemy H0)

#warto zrobic hist(dane$dochod)

mean(dane$dochod)

#H0 srednia wartosc zmiennej dochod jest mniejsza od 68400

t.test(dane$dochod, mu= 68400, alternative='g')

#p-value = 0.9574 > 0.05

#Przyjmujemy hipoteze H0