Zaliczenie przedmiotu: Matematyka stosowana ze statystyką 2

1. Laboratorium 1

Struktury danych w R

Obiekt	Typy	Różne typy
Wektor (*vector*)	numeryczny, znakowy, zespolony, logiczny	Nie
Macierz (*matrix*)	numeryczny, znakowy, zespolony, logiczny	Nie
Tabela (*table*)	numeryczny, znakowy, zespolony, logiczny	Nie
Ramka danych (*data.frame*)	numeryczny, znakowy, zespolony, logiczny	Tak
Lista (*list*)	numeryczny, znakowy, zespolony, logiczny, funkcja, wyrażenie	Tak

Operatory:

- operatory arytmetyczne: +, -, *, /, ^, %% (dodawanie, odejmowanie, mnożenie, dzielenie, potęgowanie, dzielenie modulo);
- operatory logiczne: !, |, & (negacja, alternatywa, koniunkcja);
- operatory relacyjne: >, <, >=, <=, ==, != .

Pełna lista operatorów na stronie:

https://cran.r-project.org/doc/manuals/r-release/R-lang.html#Operators

Wybrane (wbudowane) funkcje

Funkcja	Opis
log(x)	Logarytm naturalny z x
exp(x)	Liczba e podniesiona do potęgi x
log(x, n)	Logarytm z x przy podstawie n
sqrt(x)	Pierwiastek kwadratowy z x
factorial(n)	$n! = 1 \cdot 2 \cdot \ldots \cdot n$
choose(n, k)	Symbol Newtona
abs(x)	Wartość bezwzględna z x
round(x, digits=n)	Zaokrągla x do n miejsc po przecinku
sin(x)	Sinus x (x w rad)
cos(x)	Cosinus x (x w rad)
tan(x)	Tangens x (x w rad)

Wektory w R możemy tworzyć wykorzystując:

- funkcję c(), (*combine* - połącz, np. c(1, 3, 5, 9, 10));
- operator : tworzący ciąg arytmetyczny o różnicy 1 (np. 1:10);
- funkcję seq(), (*sequence* - sekwencja) tworzy ciąg arytmetyczny (np. seq(−5, 5, by = 0.2))
- funkcję rep(), (*replicate* - powtórz) powtarza elementy zadaną liczbę razy (np. rep(c(1,2), 2)).

Funkcja	Opis
length(**x**)	Długość (liczba elementów) wektora **x**
max(**x**)	Największa wartość wśród elementów wektora **x**
min(**x**)	Najmniejsza wartość wśród elementów wektora **x**
sum(**x**)	Suma wszystkich wartości wektora **x**
prod(**x**)	Iloczyn wszystkich wartości wektora **x**
sort(**x**, decreasing = FALSE)	Sortuje (rosnąco) wartości wektora **x**; gdy TRUE - malejąco
sample(**x**, *n*, replace=TRUE)	Losowanie *n* elementów spośród wartości wektora **x** ze zwracaniem (replace =TRUE) lub bez (replace=FALSE)
which(**x**)	Zwraca te indeksy wektora logicznego **x**, które mają wartość TRUE, np which(**x** == 5) podaje indeksy elementów wektora **x** równych 5
unique(**x**)	Usuwa duplikaty elementów wektora **x**
x [*i*]	Indeksowanie wektora - odwołanie się do elementu wektora o indeksie *i*
summary(**x**)	Podsumowanie dla wektora **x** (minimum, maksimum, średnia, kwantyle) Dla wektora elementów kategorycznych (*factor*) zlicza elementy na poszczególnych poziomach

Macierze

- matrix(*dane*, *r*, *k*) - tworzy macierz o *r* wierszach i *k* kolumnach z wektora *dane*
- cbind($\mathbf{x}_1$, …, $\mathbf{x}_n$) - tworzy macierz łącząc podane wektory jako kolumny
- rbind($\mathbf{x}_1$, …, $\mathbf{x}_n$) - tworzy macierz łącząc podane wektory jako wiersze

Funkcja	Opis
dim(**A**)	Wymiar macierzy **A**
t(**A**)	Transpozycja macierzy **A**
sum(**A**)	Suma elementów macierzy **A**
det(**A**)	Wyznacznik macierzy **A**
A %*% **B**	Mnożenie macierzy
solve(**A**)	Tworzenie macierzy odwrotnej
solve(**A**, **B**)	Rozwiązanie układu macierzowego **AX** = **B**
A[1,2]	Odwołanie się do elementu w pierwszym wierszu i drugiej kolumnie macierzy **A**

LABORATORIUM 1.

WPROWADZENIE DO ŚRODOWISKA R

WYMAGANIA: ŚRODOWISKO R, RSTUDIO, BIBLIOTEKI: dplyr, ggplot2, moments, readxl (https://cran.r-project.org)

Utworzyć nowy projekt pod nazwą Lab_R oraz skrypt pod nazwą Lab_1.

ZAD.1. Obliczyć wartości następujących wyrażeń w środowisku R:

a) $4 \cdot 5^2 + \log_3 30$ b) $\sqrt[5]{7}$ c) $\sqrt[3]{\sqrt[7]{6}}$

ZAD.2. Dane są macierze $\mathbf{A} = \begin{bmatrix} 3 & 2 & -1 \\ 4 & 1 & 0 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 7 & 2 & 12 \\ -11 & -6 & 1 \\ 3 & -2 & 3 \end{bmatrix}$. Zdefiniować macierze i obliczyć tam, gdzie to możliwe

a) wyznaczniki macierzy,
b) macierze odwrotne,
c) macierze transponowane i ich wyznaczniki,
d) iloczyny macierzy (także przez siebie),
e) iloczyn skalarny między pierwszym wierszem macierzy **A** a drugą kolumną macierzy **B** (odwołania np. **A**[1,3]).

ZAD.3. Wykorzystując zapis macierzowy rozwiązać układ równań (użyć *solve*)

$$\begin{cases} 2x + 4y - 3z + t = 71 \\ 3x - 2y + 8z - 11t = -20 \\ x + 3y + 2z + 5t = 26 \\ 4x - 3y - 5z - 3t = 49 \end{cases}$$

ZAD.4. Utworzyć wektor kwadratów liczb od 1 do 80, a następnie ustalić, które cyfry oraz jak często występują na pozycji jedności w wyznaczonych kwadratach (użyć operatora modulo oraz funkcji *summary* i *factor*).

ZAD.5. Utworzyć tablice trygonometryczne, w których zebrane będą informacje o wartościach funkcji sinus, cosinus, tangens i cotangens dla kątów od 30° do 60° co 5° (funkcje trygonometryczne w R przyjmują argumenty w radianach). W tym celu napisać funkcję **rad** (użyć *function*) do zamiany stopni na radiany (stała π w R ma nazwę *pi*), utworzyć wektor argumentów w radianach oraz ramkę danych **Tablice** (użyć *data.frame*).

ZAD.6. Utworzyć wektor 40 łańcuchów znaków następującej postaci: litera.liczba, gdzie litera to trzy duże litery X, Y, Z występujące cyklicznie, a liczba to kolejne liczby od 1 do 40 czyli X.1 Y.2 Z.3 X.4 itd. Wykorzystać funkcję *paste*, która łączy napisy.

```r
#ZADANIE 1----------------------------------------------------------------------------------------------------
#a)
4*5^2+log(30,3)
#b)
7^(1/5)
#c)
(6^(1/7))^(1/3)
#ZADANIE 2----------------------------------------------------------------------------------------------------
#tworzenie macierzy na 3 sposoby
#pierwszy sposób – łączenie wektorów w wiersze i potem w macierz
w1=c(3,2,-1)
w2=c(4,1,0)
A1=rbind(w1,w2)
#drugi sposób – łączenie wektorów w kolumny i potem macierz
k1=c(3,4)
k2=c(2,1)
k3=c(-1,0)
A2=cbind(k1,k2,k3)
#trzeci sposób - funkcja wbudowana - uwaga idziemy kolumnami!!!
A=matrix(c(3,4,2,1,-1,0),2,3)
B=matrix(c(7,-11,3,2,-6,-2,12,1,3),3,3)
#a) wyznaczniki macierzy (musi być macierz kwadratowa)
det(B)
#b) macierze odwrotne (musi być kwadratowa)
solve(B)
#c) macierze transponowane i ich wyznaczniki
t(A)
t(B)
det(t(B))
#d) iloczyn macierzy (AxA, AxB, BxA i BxB)
#A%*%A - nie można, bo jest 2x3 i 2x3, wiec środkowe nie są takie same
A%*%B
#B%*%A - nie można, bo jest 3x3 i 2x3, wiec środkowe nie są takie same
B%*%B
#e) iloczyn skalarny miedzy pierwszym wierszem macierzy A, a druga kolumna macierzy B
A[1,]%*%B[,2]
#ZADANIE 3----------------------------------------------------------------------------------------------------
M=matrix(c(2,3,1,4,4,-2,3,-3,-3,8,2,-5,1,-11,5,-3),4,4)
W=matrix(c(71,-20,26,49),4,1)
solve(M,W)
#ZADANIE 4----------------------------------------------------------------------------------------------------
#funkcja factor
wk=(1:80)^2
```

```
jed=wk%%10
x=factor(jed)
wynik=summary(x)
```

#ZADANIE 5--

```
rad=function(deg){
  deg*pi/180
}

stopnie=seq(30,60,by=5)
tablica=data.frame(
  stopnie,
  sinus=sin(rad(stopnie)),
  cosinus=cos(rad(stopnie)),
  tangens=tan(rad(stopnie)),
  cotangens=1/tan(rad(stopnie))
  )
```

#ZADANIE 6--

```
litery=rep("X","Y","Z"),length.out=40)
numery=(1:40)
lanuch=paste(litery,numery, sep="."))
print(noquote(lancuch))
```

2. Laboratorium 2

OZNACZENIA WYBRANYCH ROZKŁADÓW TYPU SKOKOWEGO W R:
binom (dwumianowy), geom (geometryczny), pois (Poissona)

PREFIKSY: ***d*** - funkcja rozkładu, ***p*** – wartość dystrybuanty, ***q*** – wartość kwantyla, ***r*** – generator liczb losowych
np. w rozkładzie dwumianowym z parametrami n i p

$$P(X=a)=\mathbf{dbinom}(a,n,p)$$

$$P(X\le a)=F(a)=\mathbf{pbinom}(a,n,p)$$

$$P(X>a)=1-\mathbf{pbinom}(a,n,p)=\mathbf{pbinom}(a,n,p,\mathbf{lower.tail=F})$$

$$F(a)=b\Rightarrow a=\mathbf{qbinom}(b,n,p)$$

WYBRANE ROZKŁADY DLA ZMIENNEJ LOSOWEJ TYPU SKOKOWEGO

- Rozkład dwumianowy z parametrami $n=1,2,...,\ p\in(0,1)$.

$$P(X=k)=\binom{n}{k}p^k q^{n-k},\ k=0,1,2,...,n,\ q=1-p$$

  Realizacja: liczba sukcesów w schemacie Bernoulli'ego

- Rozkład geometryczny z parametrem $p\in(0,1)$.

$$P(X=k)=q^k p,\ q=1-p,\ k=0,1,2,...$$

  Realizacja: **liczba porażek** do momentu wystąpienia pierwszego sukcesu

- Rozkład Poissona z parametrem $\lambda>0$.

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\ k=0,1,2,....$$

  Uwaga. Jeśli $n\ge 50,\ p\le 0.1$ i $n\cdot p\le 10$, to do celów praktycznych można przybliżać rozkład dwumianowy rozkładem Poissona:

$$\binom{n}{k}p^k q^{n-k}\approx\frac{\lambda^k}{k!}e^{-\lambda},\ \lambda=n\cdot p,$$

LABORATORIUM 2.

ZMIENNE LOSOWE TYPU SKOKOWEGO

WYBRANE ROZKŁADY TYPU SKOKOWEGO W R: binom, geom, pois

PREFIKSY: ***d*** – funkcja rozkładu, ***p*** – wartość dystrybuanty, ***q*** – wartość kwantyla, ***r*** – generator liczb losowych (np. **pbinom(*x*, *n*, *p*)** – wartość dystrybuanty rozkładu dwumianowego z parametrami n i p w punkcie x)

ZAD.1. Prawdopodobieństwo, że przeciętny student nie zrobi pewnego zadania na kolokwium wynosi $\frac{3}{7}$. Nauczyciel wybiera przypadkowo 5 prac różnych studentów. Wyznaczyć rozkład prawdopodobieństwa liczby osób spośród wybranych, które nie zrobiły tego zadania. Sprawdzić, czy rozkład jest dobrze określony.

ZAD.2. W pewnej rodzinie dwoje spośród trojga dzieci urodziło się w środę. Jakie jest prawdopodobieństwo takiego zdarzenia?

ZAD.3. Prawdopodobieństwo awarii pewnego urządzenia podczas uruchomiania wynosi 0.003. Obliczyć prawdopodobieństwo, że pierwsza awaria zdarzy się przy szóstym włączeniu.

ZAD.4. W skład pewnej wtryskarki wchodzi 300 elementów określonego rodzaju. Prawdopodobieństwo uszkodzenia w ciągu miesiąca każdego z tych elementów wynosi 0.002 i nie zależy od stanu pozostałych. Obliczyć prawdopodobieństwo uszkodzenia w ciągu miesiąca:
a) dokładnie trzech elementów,
b) nie więcej niż trzech elementów.
W obu podpunktach obliczyć przybliżenie rozkładem Poissona.

ZAD.5. Wadliwość produkowanych w pewnej firmie kości pamięci wynosi 0.4%. Pobrano losowo do kontroli partię 600 kości pamięci. Obliczyć prawdopodobieństwo, że liczba uszkodzonych kości pamięci jest większa niż 3.

ZAD.6. Rzucamy jednocześnie trzema monetami aż wypadną trzy orły. Jakie jest prawdopodobieństwo, że będziemy musieli rzucać więcej niż 5 razy?

ZAD.7. Pewne urządzenie zawiera 650 lamp. Prawdopodobieństwo przepalenia dowolnej lampy w ciągu jednej doby pracy urządzenia jest jednakowe i wynosi 0.003. Obliczyć prawdopodobieństwo, że w ciągu jednej doby pracy urządzenia przepalą się co najmniej 2 lampy.

ZAD.8. Rzucamy jednoczenie dwiema kostkami aż na obu wypadnie co najmniej 5 oczek. Obliczyć prawdopodobieństwo, że zdarzy się to:
a) w trzecim rzucie,
b) w co najmniej drugim, ale nie później niż w siódmym rzucie.

ZAD.9. W centrali telefonicznej jest 1000 linii, które działają niezależnie od siebie. Prawdopodobieństwo tego, że linia nie jest zajęta wynosi 0.88. Obliczyć prawdopodobieństwo, że liczba zajętych linii różni się od 100 o mniej niż 15.

#ZADANIE 1--

```
# w funkcji dbinom zadajemy: ilość sukcesów, ilość prób, prawdopodobieństwo sukcesu dbinom(0,5,3/7)
```

```
dbinom(0,5,3/7)
#teraz przeprowadzimy analizę dla wszystkich możliwych sukcesów
dbinom(0:5,5,3/7)
#przedstawimy to w postaci macierzowej
rozklad=rbind(x_i=0:5, p_i=dbinom(0:5,5,3/7))
#sprawdzamy czy wszystkie prawdopodobieństwa sumują się do 1
sum(dbinom(0:5,5,3/7)) #dobrze określony, bo sumuje się do 1
```

#ZADANIE 2---

```
dbinom(2,3,1/7)   #1/7 bo dwoje dzieci urodziły się jednego dnia
```

#ZADANIE 3 (będzie na kolokwium) ---

```
#w funkcji dgeom zadajemy: ilość porażek, prawdopodobieństwo sukcesu
#P(X=5) pięć porażek bo szósty jest sukces
dgeom(5,0.003)
```

#ZADANIE 4---

```
#podpunt a - dokładnie trzech elementów (rozkład dwumianowy)
dbinom(3,300,0.002)
#podpunkt b - nie więcej niż trzech elementów (rozkład dwumianowy)
pbinom(3,300,0.002) #zsumuje nam, czyli krótko mówiąc polaczenie funkcji dbinom i sum(dbinom)
#podpunkt a zapisany w rozkładzie Poissona
dpois(3,300*0.002)
#podpunkt b zapisany w rozkładzie Poissona
ppois(3,300*0.002)
```

#ZADANIE 5---

```
#(dwie wersje, obie poprawne)
#P(X>3) = 0.2210252
sum(dbinom(4:600,600,0.004))
# 1-F(3) = 1-P(X<=3)
1-pbinom(3,600,0.004)
#lub
pbinom(3,600,0.004,lower.tail=F)
#lub
pbinom(3,600,0.004,0)
```

#ZADANIE 6---

```
#prawdopodobienstwo to 1/8, bo mamy jedna sytuacje, gdy sa 3 orły i 8 możliwych kombinacji 2^3
#ilosc porażek to 4, bo więcej niż 5 razy
1-pgeom(4,1/8) #bo interesują nas od 5 porażek wzwyż
#lub
pgeom(4,1/8,0)
```

#ZADANIE 7---

```
1-pbinom(1,650,0.003)
1-ppois(1,650*0.003)

# P(X>=2) = 0.5806874
sum(dbinom(2:650,650,0.003))
# 1-F(1) = 1-P(X<=1)
```

```
1-pbinom(1,650,0.003)
# P(X>1)
pbinom(1,650,0.003,0)
```

#ZADANIE 8---

```
#podpunkt a - w trzecim rzucie
#6*6=36, 55,56,65,55 - 4/36=1/9
dgeom(2,1/9)
#podpunkt b - w co najmniej drugim, ale nie później niż w 7, czyli od 1 do 6 porażek
sum(dgeom(1:6,1/9))
pgeom(6,1/9)-pgeom(0,1/9)
```

#ZADANIE 9---

```
#prawdopodobieństwo to 1-0.88=0.12, bo rozważamy zajęte linie
#ilość prób to 1000, bo tyle linii
#liczba sukcesów między 86 a 114
sum(dbinom(86:114,1000,0.12))
```

3. Laboratorium 3

LABORATORIUM 3.

ZMIENNE LOSOWE TYPU CIĄGŁEGO

WYBRANE ROZKŁADY TYPU CIĄGŁEGO W R: norm, t, chisq, f

PREFIKSY: ***d*** – funkcja gęstości, ***p*** – wartość dystrybuanty, ***q*** – wartość kwantyla, ***r*** – generator liczb losowych (np. **pnorm(*x*, *m*, σ)** – wartość dystrybuanty rozkładu normalnego z parametrami m i σ w punkcie x)

ZAD.1. Obliczyć kwantyle:

a) $u(0.98)$;
b) $t(0.95, 18)$;
c) $\chi^2(0.975, 23)$;
d) $F(0.99, 5, 24)$.

ZAD.2. Zmienna losowa X ma rozkład normalny $N(3,6)$. Obliczyć prawdopodobieństwo:

a) $P(X<5)$,
b) $P(X>4)$,
c) $P(-1<X\le 1)$,
d) $P(|X-4|\le 0.5)$,
e) $P(|3X-8|<1)$,
f) $P(|X+1|\ge 7)$,
g) $P(|2X-3|>4)$.

ZAD.3. Czas świecenia żarówek pochodzących z masowej produkcji jest zmienną losową X o rozkładzie normalnym N(200 h, 10 h). Oblicz, ile przeciętnie żarówek spośród 10000 świeci krócej niż 175 h.

ZAD.4. Przy założeniu, że wyniki w skoku wzwyż mężczyzn mają rozkład normalny z parametrami 2.25 m oraz 0.2 m, obliczyć:

a) ilu zawodników na 40 osiągnie w skoku wzwyż co najmniej 2.3 m,
b) jaki jest wynik uzyskany przez zawodników, poniżej którego jest 20% najsłabszych rezultatów?

ZAD.5. Przyjmując, że opóźnienie pociągu do Poznania jest zmienną losową o rozkładzie normalnym N(13 min, 18 min), obliczyć prawdopodobieństwo, że pociąg, który miał przyjechać o 14.25 przyjedzie:

c) między 14.40 a 14.45,
d) po 14.50.

ZAD.6. Zmienna losowa ma rozkład $N(25, 8)$. Wyznaczyć nieznane wartości całkowite k_1, k_2, k_3, k_4, jeżeli wiadomo, że zmienna ta przyjmuje wartość:

a) mniejszą niż k_1 z prawdopodobieństwem 0.5987,
b) większą od k_2 z prawdopodobieństwem 0.734,
c) odchylającą się od średniej nie więcej niż o k_3 z prawdopodobieństwem 0.468,
d) odchylającą się od średniej nie mniej niż o k_4 z prawdopodobieństwem 0.617.

```
#pamietać o włączeniu pakietu ggplot, bo są ładniejsze wykresy

#a) rozkład normalny N(0,1) - najważniejszy rozkład
ggplot(data.frame(x=c(-5,5)), aes(x))+
  stat_function(fun=dnorm,col="blue",size=1.25)+
  ylab("Gęstość rozkładu normalnego N(0,1)")
#u~N(0,1)

#b)
```

```r
ggplot(data.frame(x=c(-5,5)), aes(x))+
  stat_function(fun=dt,args=list(df=18),col="green",size=1.25)+
  ylab("Gęstość rozkładu t Studenta")
#df - linia stopni swobody, bierze sie z licznosci probki

#c)
ggplot(data.frame(x=c(-2,75)), aes(x))+
  stat_function(fun=dchisq,args=list(df=23),col="red",size=1.25)+
  ylab("Gęstość rozkładu chi kwadrat")

#d)
ggplot(data.frame(x=c(-2,7)), aes(x))+
  stat_function(fun=df,args=list(df1=5,df2=24),col="brown",size=1.25)+
  ylab("Gęstość rozkładu Fishera")
```

#ZADANIE 1 (na kolokwium)--

```r
#a) kwantyl rozkładu normalnego
qnorm(0.98)
#odp. 2.053749

#b) kwantyl rozkładu t Studenta
qt(0.95,18)
#odp. 1.734064

#c) kwantyl rozkładu chi kwadrat
qchisq(0.975,23)
#odp. 38.07563

#d) kwantyl rozkładu Fishera
qf(0.99,5,24)
#odp. 3.89507
```

#ZADANIE 2--

```r
#N(3,6)

#a) P(X<5)
#w kolejności 5 bierze się z P(X<5), a 3 i 6 z rozkładu N(3,6)
pnorm(5,3,6)
# F(5) = 0.6305587

#b) P(X>4)
1-pnorm(4,3,6)
# 1-F(4) = 0.4338162

#c) P(-1<X<=1)
pnorm(1,3,6)-pnorm(-1,3,6)
#P(-1<X<=1)=F(1)-F(-1)= 0.1169488

#d) P(|X-4|<=0.5)
pnorm(4.5,3,6)-pnorm(3.5,3,6)
# P(4-0.5<=X<=4+0.5)
#F(4.5)-F(3.5) = 0.06549957

#e) P(|3X-8|<1)
pnorm(3,3,6)-pnorm(7/3,3,6)
```

```
# P(8-1<3X<8+1) = P(7/3<X<3)
#F(3)-F(7/3)=0.04423588
```

```
#f) P(|X+1|>=7)
1-(pnorm(6,3,6)-pnorm(-8,3,6))
# P(X<=-1-7)+P(X>=-1+7) = F(-8)+1-F(6) = 0.341914
```

```
#g) P(|2X-3|>4)
1-(pnorm(7/2,3,6)-pnorm(-1/2,3,6))
pnorm(-1/2,3,6)+pnorm(7/2,3,6,0)
# P(2X<3-4)+P(2X>3+4) = F(-1/2)+1-F(7/2) = 0.7466277
#to zero sprawi, że policzy nam od drugiej strony
```

#ZADANIE 3 (na kolokwium)--

```
#N(200,10)
#prawdopodbieństwo dla jednej razy ilość żarówek
# P(X<175)
pnorm(175,200,10)
10000*pnorm(175,200,10)
# 62  żarówki
```

#ZADANIE 4---

```
#N(2.25,0.2)
#a) ilu zawodników na 40 osiągnie w skoku wzwyż co najmniej 2,3
(1-pnorm(2.3,2.25,0.2))*40
round((1-pnorm(2.3,2.25,0.2))*40)
#16 zawodników
```

```
#b) jaki jest wynik uzyskany przez zawodników, poniżej którego jest 20% najsłabszych rezultatów
qnorm(0.2,2.25,0.2)
#pułap 2.08
# P(X<x) = 0.2 x-kwantyl rzędu 0.2
```

#ZADANIE 5---

```
#N(13,18)
#a) miedzy 14:40 a 14:45
pnorm(15,13,18,0)-pnorm(20,13,18,0)
```

```
#b) po 14:50
pnorm(25,13,18,0)
```

4. Laboratorium 4

LABORATORIUM 6.

ZARZĄDZANIE DANYMI

RAMKA DANYCH:

wywołanie kolumny w ramce np. Ankieta$Płeć
usuwanie kolumny w ramce np. Ankieta[,–nr] lub Ankieta$Płeć = NULL
edycja ramki np. *fix*(Ankieta)
symbol braku danych – NA

WYKRES:

histogram dla zmiennej kategorycznej – *ggplot* (zbiór danych, *aes* (*x* = zmienna)) + *geom_bar* (*fill* = "kolor", *col* = "kolor") + *ylab* ("opis")

ZAD.1. Otworzyć plik z danymi w Excel

a) zmienić nagłówki kolumn na krótsze wprowadzając nazwy: Waga, Wzrost, M.zamieszkania, Sz.średnia, ECTS, Algebra, MSzS1, Narz.inż, Prog1, WdI, L.godzin, L.sys.op, System, Wiek;
b) w kolumnie Sz.średnia zamienić łańcuchy *klasa z rozszerzoną matematyką* na *RM*;
c) plik po zmianach zapisać pod nazwą Ankieta.xlsx.

ZAD.2. Wczytać w RStudio plik Ankieta.xlsx (Environment/Import Dataset/From Excel), zmieniając na *numeric* typ zmiennych mierzalnych: Waga, Wzrost, ECTS, Algebra, MSzS1, Narz.inż, Prog1, WdI, L.godzin, L.sys.op, Wiek

a) wyświetlić podsumowanie danych (użyć *summary*) przed i po faktoryzacji zmiennych niemierzalnych (użyć *factor*) – ocenić wartości skrajne zmiennych mierzalnych;
b) w ramce Ankieta utworzyć nową zmienną Średnia, zawierającą średnią ocen z kursów;
c) przenieść kolumny z ocenami z kursów do podzbioru Ankieta.kursy (użyć *subset*);
d) napisać funkcję **zakres3sigm**, która zwróci dla dowolnej zmiennej ramkę danych z nagłówkami *lewy.kres* / *prawy.kres* jako średnią minus / plus trzy odchylenia standardowe (użyć *function*, *mean*, *sd*, *data.frame*);
e) dla zmiennej Średnia wyznaczyć ewentualne dane odstające i zastąpić je symbolem braku danych, a później średnią, zaokrągloną do części dziesiętnych (przy dużej liczbie danych użyć *which*);
f) utworzyć podzbiory danych Ankieta.M i Ankieta.K dla mężczyzn i kobiet odpowiednio (użyć *filter*);
g) dla zmiennych Waga i Wzrost wyznaczyć ewentualne dane odstające dla obu płci, zastąpić je symbolem braku danych, a później średnią, zaokrągloną do części dziesiątych;
h) utworzyć nową zmienną L.g.kody, w której zostaną umieszczone 3 przedziały liczbowe odpowiadające ustalonym kategoriom: krótko, średnio, długo (użyć *cut*) i wyświetlić liczności przedziałów.

ZAD.3. Wyznaczyć histogramy dla zmiennych M.zamieszkania, Sz.średnia i System.

#ZADANIE 2 ---

#a)

```
summary(Ankieta)
Ankieta$Płeć = factor(Ankieta$Płeć)
Ankieta$M.zamieszkania = factor(Ankieta$M.zamieszkania)
Ankieta$Sz.średnia = factor(Ankieta$Sz.średnia)
Ankieta$System = factor(Ankieta$System)
summary(Ankieta)
#(usunąć pierwszy # wykonać instrukcje i ponownie postawić #)
######Ankieta=Ankieta[-32,] #usuwanie wiersza nr 32
```

#b) - tworzenie nowej kolumny Średnia i wstawianie wartości

```
Ankieta$Średnia=(Ankieta$Algebra+Ankieta$MSzS1+Ankieta$Narz.inż.+Ankieta$Prog1+Ankieta$WdI)/5
```

#c) - dzielenie tabeli na podtabelę i usuwanie kolumn ze starej

```
#Ankieta.kursy=subset(Ankieta,select=Algebra:WdI)
######Ankieta=Ankieta[,-(7:11)] #uzywamy tego raz (usuwamy kolumny)
```

#d) - pozbycie się danych odstających poprzez funkcję odejmującą 3 sigmy (3 sigmy w prawo i 3 sigmy w lewo to 99% populacji)

```
zakres3sigm=function(x){
lewy.kres=mean(x)-3*sd(x)
prawy.kres=mean(x)+3*sd(x)
data.frame(lewy.kres,prawy.kres)
}
```

#e) -

zakres3sigm(Ankieta$Średnia)
summary(Ankieta$Średnia)
#dobre dane, bo mieszczą się w przedziale

#f) - tworzenie podzbiorów danych dla kobiet i mężczyzn

Ankieta.M=subset(Ankieta,Płeć=="M")
Ankieta.K=subset(Ankieta,Płeć=="K")

#g) - ew. dane odstające dla wagi i wzrost (zakres 3sigm musi zawierac min i max)

zakres3sigm(Ankieta.K$Wzrost)
summary(Ankieta.K$Wzrost)

zakres3sigm(Ankieta.K$Waga)
summary(Ankieta.K$Waga)

zakres3sigm(Ankieta.M$Wzrost)
summary(Ankieta.M$Wzrost)

zakres3sigm(Ankieta.M$Waga)
summary(Ankieta.M$Waga)

#ręcznie zmieniamy w ankietach dane odstające na średnią

fix(Ankieta.M)
fix(Ankieta)

#h) - tworzenie nowej zmiennej i przydzielenie jako wartość przedziałów

Ankieta$L.g.kody=cut(Ankieta$L.godzin,c(0,5,10,24))
summary(Ankieta$L.g.kody)

#ZADANIE 3--

#rysowanie histogramów (pamiętać o włączeniu pakietu ggplot2)

ggplot (Ankieta, aes(M.zamieszkania)) +geom_bar (fill = "red", col = "black") + ylab ("Liczność")
ggplot (Ankieta, aes(Sz.średnia)) +geom_bar (fill = "blue", col = "black") + ylab ("Liczność")
ggplot (Ankieta, aes(System)) +geom_bar (fill = "green", col = "black") + ylab ("Liczność")

5. Laboratorium 5

LABORATORIUM 7.

ELEMENTY STATYSTYKI OPISOWEJ

RAMKA DANYCH:
wywołanie zmiennej w ramce np. Ankieta.M$Waga

NIEKTÓRE STATYSTYKI: *mean*, *quantile*(zmienna, rząd), *median*, *IQR*, *var*, *sd*, *skewness*, *kurtosis*

WYKRESY:
histogram dla zmiennej mierzalnej – *ggplot* (zbiór danych, *aes* (*x* = zmienna)) + *geom_histogram* (*fill* = "kolor", *col* = "kolor", *binwidth* = szerokość klasy) + *ylab* ("opis")
diagram łodyga i liście – *stem* (zmienna)
pudełko ramka wąsy – *ggplot* (zbiór danych, *aes* (*x* = zmienna kategoryczna, *y* = zmienna mierzalna)) + *geom_boxplot* (*fill* = "kolor", *col* = "kolor")

ZAD.1. Napisać funkcję **parametry.opisowe**, która dla dowolnej zmiennej wyznaczy parametry opisowe (do łączenia wartości użyć *rbind*):

- srednia,
- kwartyl.1,
- mediana,
- kwartyl.3,
- min,
- max,
- rozstep.empiryczny = *max*(*x*) – *min*(*x*),
- rozstep.miedzykwartylowy,
- wariancja,
- odchylenie.standardowe,
- wspolczynnik.zmiennosci = *sd*(*x*)/*mean*(*x*),
- wspolczynnik.asymetrii,
- wspolczynnik.skupienia.

ZAD.2. Dla zmiennych Waga i Wzrost w grupie mężczyzn

a) wyznaczyć i zinterpretować parametry opisowe;
b) narysować histogram (przyjąć szerokość klasy 10);
c) wyznaczyć i zinterpretować diagram łodyga i liście;
d) wyznaczyć i zinterpretować wykresy ramka-wąsy z podziałem ze względu na płeć.

ZAD.3. Dla zmiennej Wzrost w grupie mężczyzn wyznaczyć szereg rozdzielczy przedziałowy (krok – 10 cm, od minimum obciętego w dół z dokładnością do 10 cm, użyć *table* i *cut*) i utworzyć pomocnicze zmienne Wzrost.środki i Wzrost.wagi. Dla tak zgrupowanych danych obliczyć średnią i odchylenie standardowe stosując pomocniczą zmienną Wzrost.szereg (użyć *rep*).
Czy otrzymane średnia i odchylenie standardowe są takie same jak parametry dla zmiennej Wzrost w grupie mężczyzn bez grupowania?

ZAD.4. Dla zmiennej Średnia

a) wyznaczyć i zinterpretować parametry opisowe;
b) wyznaczyć i zinterpretować wykresy ramka-wąsy z podziałem ze względu na płeć, miejsce zamieszkania i szkołę średnią.

summary(Ankieta)

#ZADANIE 1 - Napisać funkcję parametry.opisowe, która dla dowolnej zmiennej wyznaczy parametry opisowe (do łączenia wartości użyć rbind):

```
parametry.opisowe=function(x){
  rbind(
    srednia=mean(x),
    kwartyl.1=quantile(x,0.25), #poniżej tej wartości jest 25% populacji
    mediana=median(x), #wartość centralna
    kwartyl.3=quantile(x,0.75), #poniżej tej wartości jest 75% populacji
    min=min(x),
    max=max(x),
    rozstep.empiryczny=max(x)-min(x), #przedział w jakim działamy
    rozstep.miedzykwartylowy=IQR(x), #odległość pomiędzy dwoma kwartylami
    wariancja=var(x), #średnie kwadratowe odchylenie od średniej, w tym zadaniu kg^2
    odchylenie.standardowe=sd(x),
    wspolczynnik.zmiennosci=sd(x)/mean(x), #pozwala zaobserwować w przypadku dwóch próbek, która jest bardziej zróżnicowana, podać w %
    wspolczynnik.asymetrii=skewness(x), #czy wartość jest dodatnia czy ujemna - prawostronna(dodatnia) asymetria oznacza, że prawy bok wykresu idzie sobie powoli, ujemna odwrotnie
    #dla skewness - trzeba włączyć w packages pakiet moments
    wspolczynnik.skupienia=kurtosis(x) #mniejsza od 3, mniej skupiony rozkład wokół średniej niż w rozkładzie normalnym
  )
}
```

#ZADANIE 2 - Dla zmiennych Waga i Wzrost w grupie mężczyzn:

#a) wyznaczyć i zinterpretować parametry opisowe;

parametry.opisowe(Ankieta.M$Waga)

#WYNIKI:
srednia - średnia waga zbadanych mężczyzn wynosi 74.01 kg
kwartyl.1 - waga 25% mężczyzn nie przekracza 64.75 kg
mediana - waga 50% mężczyzn nie przekracza 73.5 kg
kwartyl.3 - waga 75% mężczyzn nie przekracza 81 kg
min, max - waga minimalna wynosi 45 kg, a maksymalna 104 kg
rozstep.empiryczny - waga mężczyzn zmienia się w zakresie 59 kg
rozstep.miedzykwartylowy - 50% środkowych wartości wagi zmienia się w zakresie 16.25 kg
wariancja - przeciętne kwadratowe odchylenie wagi mężczyzn od średniej wynosi 151.55 kg^2
odchylenie.standardowe - przeciętne odchylenie wagi mężczyzn od średniej wynosi 12.31 kg
wspolczynnik.zmiennosci - odchylenie standardowe wagi mężczyzn stanowi 16.63 % średniej
wspolczynnik.asymetrii - rozkład wagi mężczyzn jest prawostronnie asymetryczny, gdyż wsp. as. jest dodatni
wspolczynnik.skupienia - rozkład wagi mężczyzn jest słabiej skupiony wokół średniej niż w rozkładzie normalnym, gdyż kurtoza jest mniejsza niż 3

#b) narysować histogram (przyjąć szerokość klasy 10);

hist(Ankieta.M$Waga)
ggplot (Ankieta.M, aes(Waga)) + geom_histogram (fill = "red", col = "black", binwidth = 10) + ylab ("Liczność")

#c) wyznaczyć i zinterpretować diagram łodyga i liście;

stem(Ankieta.M$Waga) #można to interpretować jak godziny przyjazdów na przystanku, czyli cyfra po lewej stronie i jedna z cyfr po |,
#np. 5 | 355589 -> 53,55,55,55,58,59

#d) wyznaczyć i zinterpretować wykresy ramka-wąsy z podziałem ze względu na płeć.

boxplot(Ankieta.M$Waga)

#dolna krawędź prostokąta to kwartyl 1, pogrubiona kreska na środku to mediana, a górna krawędź prostokąta to kwartyl 3, krawędzie kreski pionowej to max i min

ggplot (Ankieta, aes (x = Płeć, y = Waga)) + geom_boxplot (fill = c("lightpink","lightblue"), col = "black")
parametry.opisowe(Ankieta.M$Wzrost)

#WYNIKI:
srednia - średni wzrost zbadanych mężczyzn wynosi 180.54 cm
kwartyl.1 - wzrost 25% mężczyzn nie przekracza 177 cm
mediana - wzrost 50% mężczyzn nie przekracza 180 cm
kwartyl.3 - wzrost 75% mężczyzn nie przekracza 185 cm
min, max - wzrost minimalny wynosi 163 cm, a maksymalny 195 cm
rozstep.empiryczny - wzrost mężczyzn zmienia się w zakresie 32 cm
rozstep.miedzykwartylowy - 50% środkowych wartości wzrostu zmienia się w zakresie 8 cm
wariancja - przeciętne kwadratowe odchylenie wzrostu mężczyzn od średniej wynosi 37.47 cm^2
odchylenie.standardowe - przeciętne odchylenie wzrostu mężczyzn od średniej wynosi 6.12 cm
wspolczynnik.zmiennosci - odchylenie standardowe wzrostu mężczyzn stanowi 3.39 % średniej
wspolczynnik.asymetrii - rozkład wzrostu mężczyzn jest lewostronnie asymetryczny, gdyż wsp. as. jest ujemny
wspolczynnik.skupienia - rozkład wzrostu mężczyzn jest bardziej skupiony wokół średniej niż w rozkładzie normalnym, gdyż kurtoza jest większa niż 3

#b) narysować histogram (przyjąć szerokość klasy 10);

hist(Ankieta.M$Wzrost)
ggplot (Ankieta.M, aes(Wzrost)) +   geom_histogram (fill = "red", col = "black", binwidth = 10) +   ylab ("Liczność")

#c) wyznaczyć i zinterpretować diagram łodyga i liście;

stem(Ankieta.M$Wzrost) #można to interpretować jak godziny przyjazdów na przystanku, czyli cyfra po lewej stronie i jedna z cyfr po |,
#np. 16 | 3   ->   163

#d) wyznaczyć i zinterpretować wykresy ramka-wąsy z podziałem ze względu na płeć.

boxplot(Ankieta.M$Wzrost)

#dolna krawędź prostokąta to kwartyl 1, pogrubiona kreska na środku to mediana, a górna krawędź prostokąta to kwartyl 3, krawędzie kreski pionowej to max i min

ggplot (Ankieta, aes (x = Płeć, y = Wzrost)) +   geom_boxplot (fill = c("lightpink","lightblue"), col = "black")

#ZADANIE 3 - Dla zmiennej Wzrost w grupie mężczyzn wyznaczyć szereg rozdzielczy przedziałowy (krok – 10 cm, od minimum obciętego w dół z dokładnością do 10 cm, użyć table i cut) i utworzyć pomocnicze zmienne Wzrost.środki i Wzrost.wagi. Dla tak zgrupowanych danych obliczyć średnią i odchylenie standardowe stosując pomocniczą zmienną Wzrost.szereg (użyć rep). Czy otrzymane średnia i odchylenie standardowe są takie same jak parametry dla zmiennej Wzrost w grupie mężczyzn bez grupowania?

table(cut(Ankieta.M$Wzrost,c(160,170,180,190,200)))
Wzrost.środki=c(165,175,185,195) #wektor wartości środkowych
Wzrost.wagi=c(8,42,42,4) #wektor ilości elementów w danym przedziale

#czyli będziemy mieli 8 razy 165, 42 razy 175, 42 razy 185 i 4 razy 195

Wzrost.szereg = rep(Wzrost.środki,Wzrost.wagi) #tyle ile elementów w przedziale tyle nam powtarza wyników
parametry.opisowe(Wzrost.szereg)

#zamiast używać dokładnie danych, używamy szeregu (sztucznie utworzonej danej), czyli zwracamy wartości ze środka, czyli zamiast wzrostów 163,164,165,166,167 przyjmujemy 5 razy 165 (środek)

6. Laboratorium 6

LABORATORIUM 8.

TEST ZGODNOŚCI I ESTYMACJA DLA JEDNEJ POPULACJI

(w każdym teście sformułować hipotezy, podać statystykę, poziom *p* oraz wniosek)

ESTYMACJA PRZEDZIAŁOWA DLA ODCHYLENIA STANDARDOWEGO W ROZKŁADZIE NORMALNYM:

MODEL 1 $\sigma \in \left(\sqrt{ns^2 / \chi^2\left(1-\frac{\alpha}{2}, n-1\right)}, \sqrt{ns^2 / \chi^2\left(\frac{\alpha}{2}, n-1\right)} \right)$ $\qquad s^2 = \frac{n-1}{n}\hat{s}^2$

WYKRES:

krzywa gęstości – *ggplot* (*data.frame* (*x* = *c*(xmin, xmax)), *aes* (*x*)) + *stat_function* (*fun* = gęstość, *args* = stopnie swobody, *col* = "kolor", *size* = grubość linii) + *ylab* ("*Gęstość rozkładu …*")

histogram dla zmiennej mierzalnej z gęstością teoretyczną – *ggplot* (zbiór danych, *aes* (*x* = zmienna)) + *geom_histogram* (*aes* (*y* = *..density..*), *fill* = "kolor", *col* = "kolor", *binwidth* = szerokość klasy)+*stat_function* (*fun* = gęstość, *args* = *list* (*mean* (zmienna), *sd* (zmienna)), *col* = "kolor", *size* = grubość linii) + *ylab* ("opis")

WE WSZYSTKICH ZADANIACH ZAKŁADAMY, ŻE POPULACJĄ GENERALNĄ SĄ WSZYSCY STUDENCI I-GO ROKU NA WI

ZAD.1. Wyświetlić w jednym układzie współrzędnych gęstości rozkładów stosowanych w statystyce matematycznej:

a) *t* Studenta (*dt*) dla stopni swobody: 1, 2, 10, 50 (przyjąć zakres od –4 do 4),

b) χ^2 (*dchisq*) dla stopni swobody: 2, 10, 20, 50 (przyjąć zakres od –1 do 100).

Wyciągnąć wnioski odnośnie rozkładów granicznych.

ZAD.2. Dla zmiennej Wzrost w grupie mężczyzn

a) dokonać wstępnej oceny zgodności z rozkładem normalnym w populacji generalnej na podstawie histogramu z gęstością teoretyczną (*dnorm*);

b) na poziomie istotności 0.05 sprawdzić założenie o normalności rozkładu testem Shapiro-Wilka (użyć *shapiro.test*);

c) wyznaczyć przedziały ufności dla średniej wzrostu w populacji generalnej (poziomy ufności 0.95 oraz 0.98, użyć *t.test*) – jak poziom ufności wpływa na szerokość przedziału ufności?

d) napisać funkcję **przedzial.odchylenie**, która zwróci wartości *ocena.dolna* i *ocena.gorna* zgodne z modelem 1 estymacji odchylenia standardowego (funkcja *var* oblicza wariancję $\hat{s}^2$);

e) wyznaczyć przedział ufności dla odchylenia standardowego wzrostu w populacji generalnej (poziom ufności 0.97);

ZAD.3. Na poziomie istotności 0.01 sprawdzić, czy można szacować średnią i odchylenie standardowe dla liczby godzin spędzanych przy komputerze w ciągu doby w populacji generalnej.

#ZADANIE 1---

#a)

```
ggplot (data.frame (x = c(-4, 4)), aes (x)) +
  stat_function (fun = dt, args = 1, col = "blue", size = 1.25) +
  stat_function (fun = dt, args = 2, col = "red", size = 1.25) +
  stat_function (fun = dt, args = 10, col = "green", size = 1.25) +
  stat_function (fun = dt, args = 50, col = "brown", size = 1.25) +
  stat_function (fun=dnorm,col="black", size=1.25)+
  ylab ("Gęstość rozkładu t Studenta")
```

#im więcej stopni swobody tym wykres bardziej podobny do wykresu rozkładu normalnego (czyli można stosować rozkład normalny zamiast t-studenta dla dostatecznie dużej próby)
#dla dostatecznie dużej próby rozkład t-studenta jest zbieżny do rozkładu normalnego
#OD PANA:
Rozkład t Studenta jest zbieżny do rozkładu normalnego standaryzowanego,
więc rozkład t Studenta można zastąpić rozkładem normalnym dla dużej próby

#b)

```
ggplot (data.frame (x = c(-1, 100)), aes (x)) +
  stat_function (fun = dchisq, args = 2, col = "blue", size = 1.25) +
  stat_function (fun = dchisq, args = 10, col = "red", size = 1.25) +
  stat_function (fun = dchisq, args = 20, col = "green", size = 1.25) +
  stat_function (fun = dchisq, args = 50, col = "brown", size = 1.25) +
  ylab ("Gęstość rozkładu chi^2")
```

#dla dostatecznie dużej próby rozkład chi^2 jest zbieżny do rozkładu normalnego

#OD PANA:
Rozkład chi kwadrat jest zbieżny do rozkładu normalnego,
więc rozkład chi kwadrat można zastąpić rozkładem normalnym dla dużej próby

#ZADANIE 2---

#a)

```
ggplot (Ankieta.M, aes(Wzrost)) +
  geom_histogram (aes (y = ..density..), fill = "blue", col = "black", binwidth = 5) +
  stat_function (fun = dnorm, args = list (mean (Ankieta.M$Wzrost), sd (Ankieta.M$Wzrost)), col = "red", size =
1.25)+
  ylab ("Częstość")
```

#na pierwszy rzut oka nie zgadza się z rokładem normalnym
#OD PANA:
Przy uogólnianiu wyników badania próbnego na pop.gen,
estymacja i testowanie parametrów wymaga zgodności z rozkładem normalnym

#H0 - hipoteza zerowa, którą testujemy
#H1 - hipoteza alternatywna dla tej, którą testujemy
#p-value - wartość na podstawie, której będziemy podejmować decyzję

#OD PANA:
Struktura testu w skrypcie:
H0:
H1:
instrukcja w R (bez #)
Statystyka testowa: =
p-value =
Wniosek:

#b) na poziomie istotności 0.05 sprawdzić założenie o normalności rozkładu
#testem ShapiroWilka (użyć shapiro.test);

#---
# Struktura testu w skrypcie:
# Hipoteza zerowa H0: rozkład wzrostu mężczyzn w pop.gen jest normalny
# Hipoteza alternatywna H1: ~H0 (lub pełnym zdaniem)
shapiro.test(Ankieta.M$Wzrost)
# Statystyka testowa: W = 0.98665
# p-value = 0.44
# Wniosek: alfa = 0.05 < p, więc nie ma podstaw do odrzucenia hipotezy H0,
# tzn. na poziomie istotności 0.05 różnice między rozkładem normalnym
# a rozkładem empirycznym nie są statystycznie istotne
#---

#OD PANA:
Poziom istotności (alfa) to prawdopodobieństwo popełnienia błędu w teście,
polegającego na odrzuceniu hipotezy H0, gdy jest ona prawdziwa
alfa = 0.05
p-value to najmniejszy poziom istotności pozwalający odrzucić hipotezę H0,
wyznaczany na podstawie statystyki testowej
Decyzje w pakietach statystycznych:
jeśli alfa >= p, to odrzucamy hipotezę H0 na korzyść hipotezy alternatywnej,
jeśli alfa < p, to nie ma podstaw do odrzucenia hipotezy H0

Można wyznaczać przedziały ufności i weryfikować hipotezy parametryczne
dla wzrostu mężczyzn w pop.gen

#c) wyznaczyć przedziały ufności dla średniej wzrostu w populacji generalnej

#(poziomy ufności 0.95 oraz 0.98, użyć t.test) – jak poziom ufności wpływa na szerokość przedziału ufności?

```
t.test(Ankieta.M$Wzrost,conf.level = 0.95 )
# Przedział liczbowy (179.3013, 181.7820) z prawdopodobieństwem 0.95
# obejmuje prawdziwą nieznaną średnią wzrostu mężczyzn w pop.gen
```

```
t.test(Ankieta.M$Wzrost,conf.level = 0.98 )
# Przedział liczbowy (179.0633, 182.0200) z prawdopodobieństwem 0.98
# obejmuje prawdziwą nieznaną średnią wzrostu mężczyzn w pop.gen
```

```
t.test(Ankieta.M$Wzrost,conf.level = 1 ) #czyli nie otrzymamy 100% pewności, bo otrrzymamy od minus niesk. do niesk.
# Im większy poziom ufności tym szerszy przedział ufności
```

#d) napisać funkcję przedzial.odchylenie, która zwróci wartości ocena.dolna i ocena.gorna zgodne z modelem 1 estymacji odchylenia standardowego (funkcja var oblicza wariancję (s^)^2);

```
przedzial.odchylenie=function(x,ufnosc){
  n=length(x)
  alfa = 1-ufnosc
  kwantyl.1 = qchisq(1-alfa/2,n-1)
  kwantyl.2 = qchisq(alfa/2,n-1)
  licznik = (n-1)*var(x)
  data.frame(
    ocena.dolna = sqrt(licznik/kwantyl.1),
    ocena.gorna = sqrt(licznik/kwantyl.2)
  )
}
```

#e) wyznaczyć przedział ufności dla odchylenia standardowego wzrostu w populacji generalnej (poziom ufności 0.97);

```
przedzial.odchylenie(Ankieta.M$Wzrost,0.97)
```

```
# Przedział liczbowy (5.287252, 7.256975) na poziomie ufności 0.97
# obejmuje prawdziwe nieznane odchylenie standardowe wzrostu mężczyzn w pop.gen
```

#ZADANIE 3 Na poziomie istotności 0.01 sprawdzić, czy można szacować średnią i odchylenie standardowe dla liczby godzin spędzanych przy komputerze w ciągu doby w populacji generalnej.

```
# Estymację stosujemy, gdy w zadaniu występuje poziom ufności (blisko 1)
# Testy stosujemy, gdy w zadaniu występuje poziom istotności (blisko 0)
```

```
#------------------------------------------------------------------------
#H0: rozkład liczby godzin spedzonych przy komputerze w ciagu doby w pop.gen jest normalny
#H1: ~H0
shapiro.test(Ankieta$L.godzin)
#Statystyka testowa: W=0.94768
#p-value = 0.0001467
#Wniosek: odrzucamy hipotezę H0 na rzecz hipotezy alternatywnej
# tzn. na poziomie istotności 0.01 różnice między rozkładem normalnym
# a rozkładem empirycznym są statystycznie istotne
# Nie można szacować średniej i odchylenia standardowego
#------------------------------------------------------------------------
```

#dodatkowo

```
ggplot (Ankieta, aes(L.godzin)) +
  geom_histogram (aes (y = ..density..), fill = "blue", col = "black", binwidth = 3) +
  stat_function (fun = dnorm, args = list (mean (Ankieta$L.godzin), sd (Ankieta$L.godzin)), col = "red", size =
1.25)+
  ylab ("Częstość")
```

7. Laboratorium 7

LABORATORIUM 9.

ESTYMACJA I TESTY DLA JEDNEJ POPULACJI

(w każdym teście sformułować hipotezy, podać statystykę, poziom *p* oraz wniosek)

WE WSZYSTKICH ZADANIACH ZAKŁADAMY, ŻE POPULACJĄ GENERALNĄ SĄ WSZYSCY STUDENCI I-GO ROKU NA WI

ZAD.1. Wiedząc, że zmienna Wzrost w grupie mężczyzn ma rozkład normalny, na poziomie istotności 0.05 zweryfikować hipotezę, że średni wzrost mężczyzn w populacji generalnej jest większy niż 179 cm (użyć *t.test*).

ZAD.2. Dla zmiennej Waga w grupie mężczyzn

a) na poziomie istotności 0.01 sprawdzić założenie niezbędne do wykonania punktów b-d;
b) wyznaczyć przedział ufności dla średniej wagi w populacji generalnej (poziom ufności 0.96);
c) wyznaczyć przedział ufności dla odchylenia standardowego wagi w populacji generalnej (poziom ufności 0.99);
d) na poziomie istotności 0.05 zweryfikować hipotezę, że średnia waga w populacji generalnej wynosi 77 kg.

ZAD.3. Wylosowaną grupę 10 osób chorujących na nadciśnienie tętnicze poddano dwukrotnemu pomiarowi ciśnienia krwi przed podaniem i po podaniu pewnego leku, testowanego pod kątem skuteczności obniżania ciśnienia. Wartości ciśnienia skurczowego zawiera tabela

Przed podaniem leku	158	160	155	170	166	173	167	180	168	173
Po podaniu leku	140	155	150	167	170	162	157	163	158	175

Na poziomie istotności 0.05 sprawdzić, czy leczenie jest skuteczne (utworzyć pomocnicze zmienne Przed i Po, sprawdzić niezbędne założenie).

ZAD.4. Używając zmiennej M.zamieszkania oraz funkcji *table* i *prop.test*

a) wyznaczyć przedział ufności dla odsetka studentów mieszkających na stancji w populacji generalnej (poziom ufności 0.96);
b) na poziomie istotności 0.05 zweryfikować hipotezę, że w akademiku mieszka 40 % studentów.

ZAD.5. W populacji generalnej

a) wyznaczyć przedział ufności dla odsetka absolwentów Liceum Ogólnokształcącego (RM) (poziom ufności 0.97);
b) na poziomie istotności 0.01 zweryfikować hipotezę, że Technikum Informatyczne ukończyło mniej niż 35% studentów.

#ZADANIE 1 - Wiedząc, że zmienna Wzrost w grupie mężczyzn ma rozkład normalny, na poziomie istotności 0.05 zweryfikować hipotezę, że średni wzrost mężczyzn w populacji generalnej jest większy niż 179 cm (użyć t.test).

```
#musimy mieć info o rozkładzie normalnym! inaczej trzeba zweryfikować

# Prawdziwą nieznaną średnią w populacji generalnej oznaczamy literą m

# Test dla średniej
#H0: m = 179 (zawsze równość !!!!)
#H1: m > 179
t.test(Ankieta.M$Wzrost, alternative="greater", mu=179)
# Statystyka testowa: t = 2.4676
# stopnie swobody: 95
#p-value = 0.007696
# Wniosek: alfa = 0.05 >= p, więc odrzucamy hipotezę H0 na korzyść hipotezy alternatywnej,
# tzn. na poziomie istotności 0.05 prawdziwa średnia wzrostu mężczyzn w pop.gen
# jest istotnie większa od 179 cm
```

#ZADANIE 2 - Dla zmiennej Waga w grupie mężczyzn

```
#a) na poziomie istotności 0.01 sprawdzić założenie niezbędne do wykonania punktów b-d;

# Test zgodności Shapiro-Wilka

# H0: rozkład wagi mężczyzn w pop.gen jest normalny
# H1: ~H0
shapiro.test(Ankieta.M$Waga)
# Statystyka testowa: W = 0.97607
# p-value = 0.07602
# Wniosek: alfa=0.01<p.value więc nie ma podstaw do odrzucenia hipotezy H0
# tzn. na poziomie istotności 0.01 różnice między rozkładem normalnym
# a rozkładem empirycznym nie są statystycznie istotne
# Można szacować średnią i odchylenie standardowe, założenie można uznać za spełnione

ggplot (Ankieta.M, aes(Waga)) +
  geom_histogram (aes (y = ..density..), fill = "blue", col = "black", binwidth = 10) +
  stat_function (fun = dnorm, args = list (mean (Ankieta.M$Waga), sd (Ankieta.M$Waga)), col = "red", size = 1.25)+
  ylab ("Częstość")

#b) wyznaczyć przedział ufności dla średniej wagi w populacji generalnej (poziom ufności 0.96);
t.test(Ankieta.M$Waga, conf.level=0.96)
# Przedział liczbowy (71.39718, 76.62991) z prawdopodobieństwem 0.96
# obejmuje prawdziwą nieznaną średnią wagę mężczyzn w pop.gen

#c) wyznaczyć przedział ufności dla odchylenia standardowego wagi w populacji generalnej (poziom ufności 0.99);
przedzial.odchylenie(Ankieta.M$Waga, 0.99)
# Przedział liczbowy (10.35606, 15.08749) na poziomie ufności 0.99
# obejmuje prawdziwe nieznane odchylenie standardowe wagi mężczyzn w pop.gen

#d) na poziomie istotności 0.05 zweryfikować hipotezę, że średnia waga w populacji generalnej wynosi 77 kg.

# Test dla średniej
# Prawdziwą nieznaną średnią w populacji generalnej oznaczamy literą m
# H0: m = 77
# H1: m ≠ 77 lub  H1: m != 77 lub  H1: m =/= 77 lub  H1: ~H0
t.test(Ankieta.M$Waga, mu=77, alternative="two.sided")
t.test(Ankieta.M$Waga, mu=77)   # alternatywa
# Statystyka testowa: t = -2.3769
# p-value = 0.01946
# Wniosek: alfa = 0.05 >= p, więc odrzucamy hipotezę H0 na korzyść hipotezy alternatywnej,
# tzn. na poziomie istotności 0.05 prawdziwa średnia waga mężczyzn w pop.gen
# różni się istotnie od 77 kg
```

#ZADANIE 3 - Wylosowaną grupę 10 osób chorujących na nadciśnienie tętnicze poddano dwukrotnemu pomiarowi ciśnienia krwi przed podaniem i po podaniu pewnego leku, testowanego pod kątem skuteczności obniżania ciśnienia.

```
#Wartości ciśnienia skurczowego zawiera tabela:
#      Po podaniu leku         158 160 155 170 166 173 167 180 168 173
#      Przed podaniem leku  140 155 150 167 170 162 157 163 158 175

#Na poziomie istotności 0.05 sprawdzić, czy leczenie jest skuteczne (utworzyć pomocnicze zmienne Przed i Po,
sprawdzić niezbędne założenie).

# Próby zależne
Przed=c(158, 160, 155, 170, 166, 173, 167, 180, 168, 173)
```

Po=c(140, 155, 150, 167, 170, 162, 157, 163, 158, 175)

# Test zgodności Shapiro-Wilka (wersja skrócona)
shapiro.test(Przed-Po)
# Rozkład różnic ciśnienia nie różni się istotnie od rozkładu normalnego

# Test dla dwóch średnich w próbach zależnych
# H0: m.Przed-m.Po = 0
# H1: m.Przed-m.Po > 0 (bo leczenie ma być skuteczne)
t.test(Przed-Po, alternative = "greater")
# Statystyka testowa: t = 3.1607
# p-value = 0.005769
# Wniosek: alfa = 0.05 >= p, więc odrzucamy hipotezę H0 na korzyść hipotezy alternatywnej,
# tzn. na poziomie istotności 0.05 w pop.gen średnia ciśnienia tętniczego przed podaniem
# leku jest istotnie większa niż po podaniu leku (leczenie jest skuteczne)

#ZADANIE 4 - Używając zmiennej M.zamieszkania oraz funkcji table i prop.test

table(Ankieta$M.zamieszkania)\

#a) wyznaczyć przedział ufności dla odsetka studentów mieszkających na stancji w populacji generalnej (poziom ufności 0.96);
# Estymacja wskaźnika struktury (odsetka)
prop.test(21,120,conf.level = 0.96)
# Przedział liczbowy (0.1117629, 0.2615617) z prawdopodobieństwem 0.96
# obejmuje prawdziwy nieznany odsetek liczby osób w pop.gen. mieszkający na stancji

#b) na poziomie istotności 0.05 zweryfikować hipotezę, że w akademiku mieszka 40 % studentów.
# Prawdziwy nieznany odsetek w pop.gen oznaczamy literą p
# Test dla wskaźnika struktury
# H0: p = 0.4
# H1: ~H0
prop.test(43,120,p=0.4)
# Statystyka testowa: chi^2 = 0.70312
# p-value = 0.4017
# Wniosek: alfa = 0.05 < p, więc nie ma podstaw do odrzucenia hipotezy H0,
# tzn. na poziomie istotności 0.05 w pop.gen odsetek studentów mieszkających w akademiku
# nie różni się istotnie od 0.4

#ZADANIE 5 - W populacji generalnej

table(Ankieta$Sz.średnia)

#a) wyznaczyć przedział ufności dla odsetka absolwentów Liceum Ogólnokształcącego (RM) (poziom ufności 0.97);

Estymacja wskaźnika struktury (odsetka)

prop.test(51,120,conf.level = 0.97)

Przedział liczbowy (0.3278395, 0.5280612) z prawdopodobieństwem 0.97
obejmuje prawdziwy nieznany odsetek liczby absolwentów Liceum Ogólnokształcącego (RM)

#b) na poziomie istotności 0.01 zweryfikować hipotezę, że Technikum Informatyczne ukończyło mniej niż 35% studentów.

Test dla wskaźnika struktury
H0: p = 0.35
H1: p < 0.35

prop.test(39,120,p=0.35,alternative="less")
# Statystyka testowa: chi^2 = 0.22894
# p-value = 0.3162
# Wniosek: alfa = 0.01 < p = 0.3162, więc nie ma podstaw do odrzucenia hipotezy H0,
# tzn. na poziomie istotności 0.01 w pop.gen odsetek absolwentów Technikum Informatycznego
# nie różni się istotnie od 0.35

#domyślnie jest w prop.test alternative="two-sided", czyli różne od

8. Laboratorium 8

LABORATORIUM 10.

TESTY DLA DWÓCH POPULACJI

(w każdym teście sformułować hipotezy, podać statystykę, poziom *p* oraz wniosek)

TESTY:

zgodności (rozkład normalny) – *by* (zmienna mierzalna, zmienna grupująca, *shapiro.test*)
parametryczne – test (*data* = zbiór danych, zmienna mierzalna~grupująca)
lub test (zmienna mierzalna~grupująca, zbiór danych)

POPULACJA GENERALNA TO WSZYSCY STUDENCI I-GO ROKU NA WI

ZAD.1. Używając zmiennej Waga porównać grupę kobiet i mężczyzn w populacji generalnej
a) na poziomie istotności 0.01 sprawdzić założenie o normalności rozkładu w obu grupach testem Shapiro-Wilka;
b) na poziomie istotności 0.05 sprawdzić jednorodność wariancji w obu grupach testem Fishera (użyć *var.test*);
c) na poziomie istotności 0.05 zweryfikować testem Studenta hipotezę, że średnia wagi w populacji generalnej jest niższa w grupie kobiet (w przypadku jednorodnych wariancji wyłączyć poprawkę Welcha: *var.equal = TRUE*).

ZAD.2. Na poziomie istotności 0.05 zweryfikować hipotezę, że średnia ocen z kursów w populacji generalnej zależy od płci (na tym samym poziomie istotności sprawdzić wszystkie niezbędne założenia).

ZAD.3. Na poziomie istotności 0.03 zweryfikować hipotezę, że mężczyźni preferujący system inny niż Windows spędzają przed komputerem więcej godzin niż pozostali (na tym samym poziomie istotności sprawdzić wszystkie niezbędne założenia).

ZAD.4. Używając zmiennych M.zamieszkania i Sz.średnia oraz funkcji *table* i *prop.test* zweryfikować hipotezę, że odsetek absolwentów LO (RM) jest większy w grupie osób mieszkających z rodziną (poziom istotności 0.05).

ZAD.5. Na poziomie istotności 0.02 zweryfikować hipotezę, że odsetek osób mieszkających w akademiku jest mniejszy w grupie kobiet.

#zerowy krok - sprawdzić czy jest rozkład normalny!!!!

#Zadanie1
Poprawne porównanie średnich w dwóch populacjach wymaga sprawdzenia założeń:
zgodności z rozkładem normalnym w obu populacjach oraz jednorodności wariancji
(założenia obejmują trzy testy istotności)

#ZADANIE 1. Używając zmiennej Waga porównać grupę kobiet i mężczyzn w populacji generalnej

#a) na poziomie istotności 0.01 sprawdzić założenie o normalności rozkładu w obu grupach testem Shapiro-Wilka;

Testy zgodności Shapiro-Wilka

#by (zmienna mierzalna, zmienna grupująca, shapiro.test) - dzięki temu nie musimy przeprowadzać całej analizy z hipotezami itp.

by(Ankieta$Waga, Ankieta$Płeć, shapiro.test)

Mężczyźni - patrz Lab_7 Zad2a

Kobiety

# H0: waga w grupie kobiet w pop. generalnej ma rozklad normalny
# H1: ~H0
shapiro.test(Ankieta.K$Waga)
# Statystyka testowa: W=0.95917
# p-value = 0.4221
# Wniosek: alfa=0.01<p.value więc nie ma podstaw do odrzucenia hipotezy H0
# tzn. na poziomie istotności 0.01 różnice między rozkładem normalnym
# a rozkładem empirycznym nie są statystycznie istotne

#b) na poziomie istotności 0.05 sprawdzić jednorodność wariancji w obu grupach testem Fishera (użyć var.test);

# Prawdziwe odchylenie standardowe w pop.gen oznaczamy σ
# Test Fishera jednorodności wariancji
# H0: σ^2.K = σ^2.M   (wariancja w grupie kobiet = wariancja w grupie mężczyzn)
# H1: ~H0
var.test(Ankieta$Waga~Ankieta$Płeć)
# Statystyka testowa: F = 0.57434
# p-value = 0.1294
# Wniosek: alfa = 0.05 < p, więc nie ma podstaw do odrzucenia hipotezy H0,
# tzn. na poziomie istotności 0.05 w pop.gen wariancje wag dla obu płci
# nie różnią się istotnie (wariancje są jednorodne)

#c) na poziomie istotności 0.05 zweryfikować testem Studenta hipotezę, że średnia wagi w populacji generalnej jest niższa w grupie kobiet (w przypadku jednorodnych wariancji wyłączyć poprawkę Welcha: var.equal = TRUE).

Test dla dwóch średnich
H0: m.K = m.M
H1: m.K < m.M
t.test(Ankieta$Waga~Ankieta$Płeć, alternative = "less", var.equal = TRUE)   #tutaj zadziałała kolejność alfabetyczna i 1. Kobiety i 2. Mężczyźni, dlatego less będzie poprawne - uważać
Statystyka testowa: t = -5.5803
p-value = 7.77e-08
Wniosek: alfa = 0.05 >= p, więc odrzucamy hipotezę H0 na korzyść hipotezy alternatywnej,
tzn. na poziomie istotności 0.05 w pop.gen średnia wagi w grupie kobiet
jest istotnie mniejsza niż średnia wagi w grupie mężczyzn

#ZADANIE 2. Na poziomie istotności 0.05 zweryfikować hipotezę, że średnia ocen z kursów w populacji generalnej zależy od płci (na tym samym poziomie istotności sprawdzić wszystkie niezbędne założenia).

#============(KROK 0)===============
Założenia:
1) Testy zgodności Shapiro-Wilka

by(Ankieta$Średnia, Ankieta$Płeć, shapiro.test)

# Mężczyźni
# H0: średnia ocen w grupie mężczyzn w pop. generalnej ma rozklad normalny
# H1: ~H0
shapiro.test(Ankieta.M$Średnia)
# Statystyka testowa: W=0.98637
# p-value = 0.4261
# Wniosek: alfa=0.05<p.value więc nie ma podstaw do odrzucenia hipotezy H0
# tzn. na poziomie istotności 0.05 różnice między rozkładem normalnym
# a rozkładem empirycznym nie są statystycznie istotne

```
# założenie można uznać za spełnione

# Kobiety
#H0: średnia ocen w grupie kobiet w pop. generalnej ma rozklad normalny
#H1: ~H0
shapiro.test(Ankieta.K$Średnia)
#Statystyka testowa: W=0.95258
#p-value = 0.3079
# Wniosek: alfa=0.05<p.value więc nie ma podstaw do odrzucenia hipotezy H0
# tzn. na poziomie istotności 0.05 różnice między rozkładem normalnym
# a rozkładem empirycznym nie są statystycznie istotne
# założenie można uznać za spełnione

#2)
# Prawdziwe odchylenie standardowe w pop.gen oznaczamy σ
# Test Fishera jednorodności wariancji
# H0: σ^2.K = σ^2.M
# H1: ~H0
var.test(Ankieta$Średnia~Ankieta$Płeć)
# Statystyka testowa: F = 1.3253
# p-value = 0.3454
# Wniosek: alfa = 0.05 < p, więc nie ma podstaw do odrzucenia hipotezy H0,
# tzn. na poziomie istotności 0.05 w pop.gen wariancje wag dla obu płci
# nie różnią się istotnie (wariancje są jednorodne)
## Założenia są spełnione, można zrobić test t bez poprawki Welcha

#3)
# Test dla dwóch średnich
# H0: m.K = m.M
# H1: ~H0
t.test(Ankieta$Średnia~Ankieta$Płeć, var.equal=TRUE)
# Statystyka testowa: t = 0.37918
# p-value = 0.7052
# Wniosek: alfa = 0.05 < p, więc nie ma podstaw do odrzucenia hipotezy H0,
# tzn. na poziomie istotności 0.05 w pop.gen średnia sredniej ocen w grupie kobiet
# nie rozni sie od tej w grupie mężczyzn
```

#ZADANIE 3. Na poziomie istotności 0.03 zweryfikować hipotezę, że mężczyźni preferujący system inny niż Windows spędzają przed komputerem więcej godzin niż pozostali (na tym samym poziomie istotności sprawdzić wszystkie niezbędne założenia).

```
## Założenia:
# 1) Testy zgodności Shapiro-Wilka

by(Ankieta.M$L.godzin, Ankieta.M$System, shapiro.test)

# Inny
#H0: Rozkład liczby godzin w grupie mężczyzn korzystających z systemu Inny w pop. generalnej ma rozklad normalny
#H1: ~H0
#Statystyka testowa: W=0.9012
#p-value = 0.1644
# Wniosek: alfa=0.03<p.value więc nie ma podstaw do odrzucenia hipotezy H0
# tzn. na poziomie istotności 0.03 różnice między rozkładem normalnym
# a rozkładem empirycznym nie są statystycznie istotne
```

```r
# założenie można uznać za spełnione

#Windows
#H0: Rozkład liczby godzin w grupie mężczyzn korzystających z systemu Windows w pop. generalnej ma rozklad normalny
#H1: ~H0
#Statystyka testowa: W=0.968
#p-value = 0.03462
# Wniosek: alfa=0.03<p.value więc nie ma podstaw do odrzucenia hipotezy H0
# tzn. na poziomie istotności 0.03 różnice między rozkładem normalnym
# a rozkładem empirycznym nie są statystycznie istotne
# założenie można uznać za spełnione

#2)
# Prawdziwe odchylenie standardowe w pop.gen oznaczamy σ
# Test Fishera jednorodności wariancji
# H0: σ^2.I = σ^2.W
# H1: ~H0
var.test(Ankieta.M$L.godzin~Ankieta.M$System)
# Statystyka testowa: F = 2.3554
# p-value = 0.02785
# Wniosek: alfa = 0.03 > p, więc odrzucamy hipotezę H0 na rzecz hipotezy alternatywnej,
# tzn. na poziomie istotności 0.03 w pop.gen wariancje liczby godzin dla obu grup
# różnią się istotnie (wariancje nie są jednorodne)
## Założenia zgodności z rozkładem normalnym są spełnione, można zrobić test t z poprawką Welcha

#3)
# Test dla dwóch średnich
# H0: m.I = m.W
# H1: m.I > m.W
t.test(Ankieta.M$L.godzin~Ankieta.M$System, alternative="greater", var.equal=FALSE)
# Statystyka testowa: t = 1.5506
# p-value = 0.0731
# Wniosek: alfa = 0.03 < p, więc nie ma podstaw do odrzucenia hipotezy H0
# tzn. na poziomie istotności 0.03 w pop.gen liczba godzin spędzanych przy komputerze wśród mężczyzn preferujących inny system operacyjny niż Windows
# nie różni się istotnie od liczby godzin przy komputerze w grupie mężczyzn preferujących Windows
```