Spis treści

Rozdział 1

Wstęp

W ramach sprawozdania z pierwszego laboratorium przeprowadzono analizę statystyczną dla przykładowych danych [1]. W rozdziale 2 przedstawiono sposób przygotowania danych wejściowych, przyjęte założenia w analizie oraz statystykę opisową. Następnie, w rozdziale 3 przedstawiono metody oraz wyniki analizy porównawczej między grupami. W rozdziale 4 przedstawiono metody oraz wyniki analizy korelacji. W rozdziale 5 podsumowano sprawozdanie.

Rozdział 2

Przygotowanie i statystyka opisowa danych

2.1 Ocena danych wejściowych

Pierwszy etap prac stanowiło załadowanie wymaganych bibliotek, pobranie przykładowych danych i wyświetlenie podstawowych informacji, co zostało przedstawione na listingu 2.1.

```
1  library("Hmisc")
2  library("dplyr")
3  library("ggpubr")
4  library("car")
5  library("dunn.test")
6  library("FSA")
7
8  df <- read.csv2("http://www.cs.put.poznan.pl/kgutowska/PSwBB/dane/
       ↪ przykladoweDane-Projekt.csv", sep = ";")
9
10 summary(df)
11 table(df$plec)
12 table(df$grupa)
```

Listing 2.1: Podstawowe operacje na danych.

Wynik polecenia *summary(df)* dla parametru wiek został przedstawiony w tabeli 2.1 natomiast dla pozostałych parametrów w tabeli 2.2.

	Wiek
Min	17.00
1st Qu.	26.50
Median	30.00
Mean	30.64
3rd Qu.	34.50
Max.	48.00

Tablica 2.1: Wyniki podstawowych statystyk dla kolumny Wiek.

	hsCRP	ERY	PLT	HGB
Min	0.3351	3.090	91.0	9.505
1st Qu.	2.2869	3.850	178.5	11.277
Median	4.1835	4.180	202.0	12.244
Mean	5.6447	4.525	220.1	12.169
3rd Qu.	7.2380	4.440	249.5	13.049
Max.	42.6499	33.000	456.0	22.232
NA's	-	-	-	2

	HCT	MCHC	MON	LEU
Min	0.0423	32.06	0.1400	4.830
1st Qu.	0.3305	34.38	0.6100	9.935
Median	0.3540	35.05	0.7400	11.400
Mean	0.3490	35.03	0.8570	11.807
3rd Qu.	0.3835	35.71	0.8975	13.775
Max.	0.4120	38.87	7.0000	17.460
NA's	-	-	1	-

Tablica 2.2: Wyniki podstawowych statystyk dla danych.

Osoby biorące udział w badaniu były osobami młodymi - średnia wieku wynosi 30 lat, a najstarszy uczestnik miał mniej niż 50 lat. Na podstawie polecenia *table* wiadomo także, że w badaniu wzięło udział 40 kobiet i 35 mężczyzn, a także, że występują dwie grupy chorych oraz jedna grupa kontrolna. W danych występują także trzy brakujące wartości. Na podstawie nazw kolumn można było wywnioskować, że dane pochodzą z badania krwi. I tak:

- hsCRP - białko C-reaktywne, bierze udział w odpowiedzi immunologicznej, wzrasta w stanach zapalnych, infekcjach oraz w wyniku zawału mięśnia sercowego [3],
- ERY - erytrocyty, odpowiadają za przenoszenie tlenu,
- PLT - płytki krwi, odgrywają istotną rolę w procesach krzepnięcia krwi,
- HGB - hemoglobina, białko, odpowiada za przenoszenie tlenu,
- HCT - hematokryt, stosunek objętości krwinek do objętości krwi [2],
- MCHC - średnie stężenie hemoglobiny w erytrocytach [4],
- MON - monocyty, odpowiadają za fagocytozę,
- LEU - leukocyty, chronią organizm przed zakażeniami i nowotworami.

W tabeli 2.3 przedstawiono oczekiwane wartości poszczególnych parametrów dla osób zdrowych, a możliwe także wyniki dla osób chorych [5]. Jest to o tyle ważne, gdyż pozwoli to na wykrycie potencjalnych błędów w danych oraz uzupełnienie brakujących wartości.

	hsCRP [mg/l]	ERY [mln/mm3]	PLT [tys/mm3]	HGB [g/dl]
Zdrowy min	0.1	3.5	150	12
Zdrowy max	3.0	5.4	400	18
Chory	powyżej 500	-	poniżej 30	-

	HCT [ml/ml]	MCHC [g/dl]	MON [10^9/l]	LEU[10^9/l]
Zdrowy min	0.361	32	0	3.5
Zdrowy max	0.503	36	0.8	9
Chory	w ciąży możliwy spadek o 30-50%	-	-	-

Tablica 2.3: Spodziewane zakresy wartości dla kolejnych parametrów.

Następny krok polegał na zidentyfikowaniu i poprawieniu wartości odstających. W tym celu porównano wartości w kolumnach z potencjalnie odstającymi wartościami (bazując na poleceniu *summary*): *hsCRP*, *HCT*, *ERY* i *MON* (pominięto wartość *NA*) z wartościami w tabeli 2.3 przy pomocy poleceń pokazanych na listingu 2.2.

```
1 sum(df$hsCRP > 10.0)
2 sum(df$HCT < 0.18)
3 sum(na.omit(df$MON) > 2.0)
4 sum(na.omit(df$ERY) > 6.0)
```

Listing 2.2: Wykrywanie wartości odstających.

Wyciągnięto następujące wnioski:

- wartość 42.6499 w kolumnie *hsCRP* jest wartością akceptowalną i prawdopodobnie nie jest błędem - w niektórych przypadkach wartości *hsCRP* przekraczają nawet 500 mg/l,
- wartość 0.0423 w kolumnie *HCT* jest prawdopodobnie błędem i zostanie poprawiona na 0.423,
- wartość 7.0 w kolumnie *MON* jest prawdopodobnie błędem i zostanie poprawiona na 0.70,
- wartość 33.0 w kolumnie *ERY* jest prawdopodobnie błędem i zostanie poprawiona na 3.3.

Polecenia pozwalające poprawić wartości odstające zostały przedstawione na listingu 2.3.

```
1 df$HCT[which(df$HCT==0.0423)] = 0.423
2 df$MON[which(df$MON==7.0)] = 0.70
3 df$ERY[which(df$ERY==33.0)] = 3.30
```

Listing 2.3: Poprawianie wartości odstających.

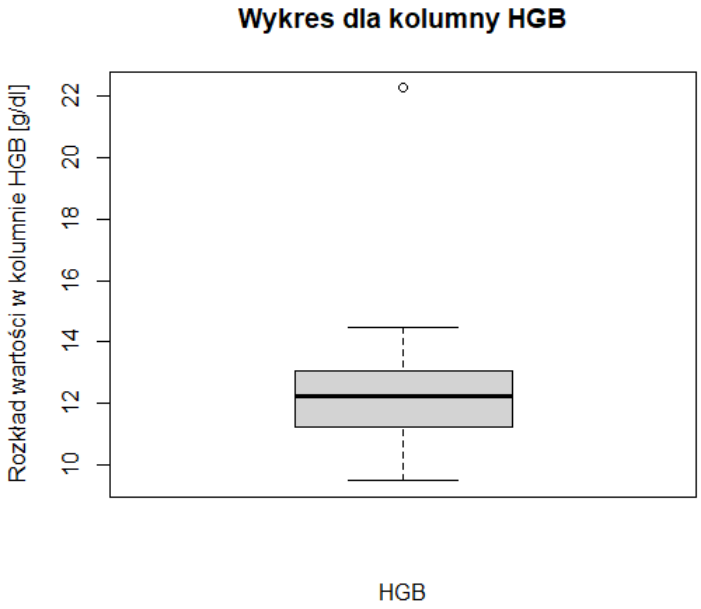
Następny krok polegał na usunięciu wartości *NA*. W tym celu użyto poleceń przedstawionych na listingu 2.4. W tym celu dla każdego parametru, w którym występowała wartość *NA* policzono średnią, grupując po grupie. Następnie wartości *NA* zostały zastąpione odpowiednimi średnimi.

```
1 df[rowSums(is.na(df)) > 0,]
2 aggregate( HGB ~ grupa, df, mean )
3 aggregate(MON ~ grupa, df, mean )
```

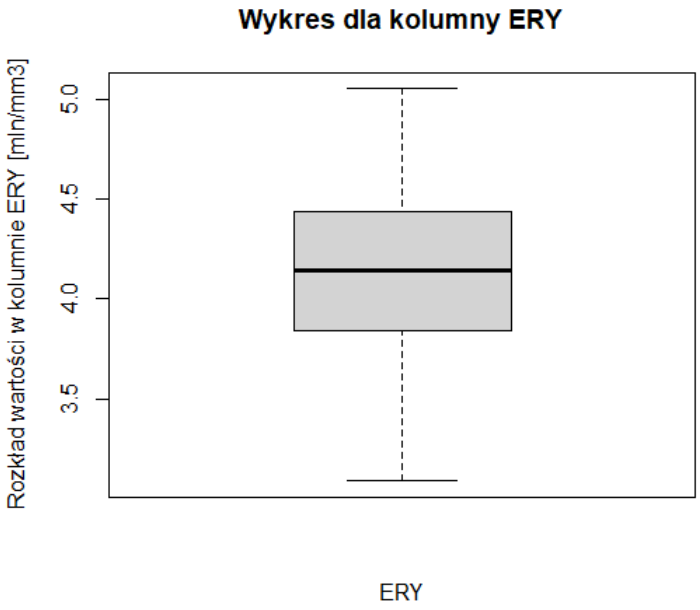
```
4
5 which(is.na(df$HGB))
6 df$HGB[13] = 12.41141
7 df$HGB[68] = 11.26357
8 which(is.na(df$MON))
9 df$MON[5] = 0.8579167
```

Listing 2.4: Poprawianie wartości *NA*.

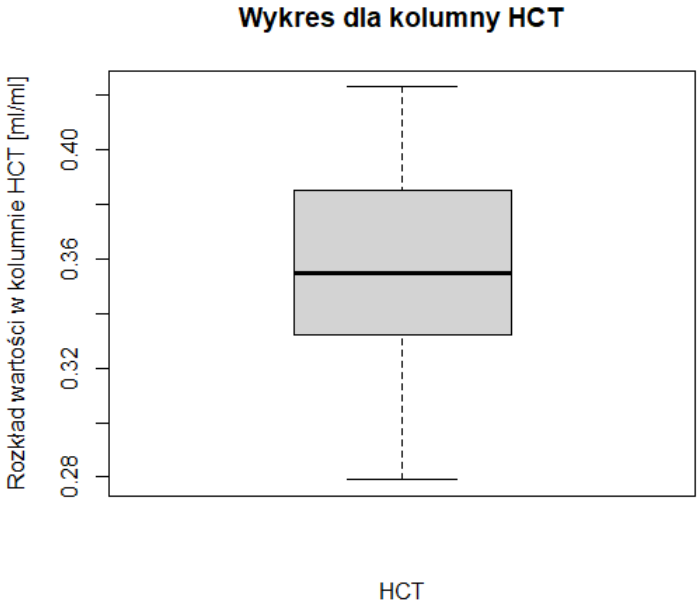
Na wykresach od 2.1 do 2.4 przedstawiono wykresy pudełkowe dla kolumn, gdzie zaszły zmiany. Można zauważyć, że zmiany wpłynęły pozytywnie na dane - otrzymane wykresy są zbliżone do modelowego wykresu pudełkowego.



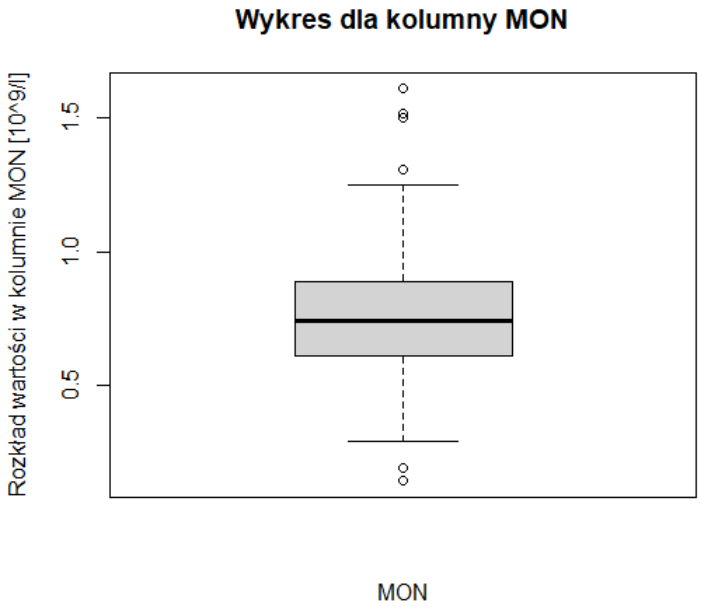
Rysunek 2.1: Wykres dla kolumny HGB.



Rysunek 2.2: Wykres dla kolumny ERY.



Rysunek 2.3: Wykres dla kolumny HCT.



Rysunek 2.4: Wykres dla kolumny MON.

2.2 Statystyki opisowe

Ze względu na fakt, że część statystyk opisowych została przedstawiona już wcześniej, na tym etapie zostaną zbadane wariancja oraz odchylenie standardowe. Dzięki pomiarowi wariancji będzie wiadomo jak duże jest zróżnicowanie pomiarów dla różnych kolumn, a dzięki odchyleniu standardowemu - jak bardzo wartości odbiegają od średniej. Wyniki zostały przedstawione w tabeli 2.4.

	Wariancja	**Odchylenie standardowe**
hsCRP	37.55244	6.128004
ERY	0.190783	0.4367871
PLT	4170.518	64.57955
HGB	3.085405	1.756532
HCT	0.001230313	0.03507581
MCHC	1.424901	1.193692
MON	0.07503833	0.2739313
LEU	7.085387	2.661839

Tablica 2.4: Otrzymane wartości wariancji i odchylenia standardowego dla poszczególnych kolumn.

Na podstawie wyników można stwierdzić, że dane dla kolumn *hsCRP* oraz dla *PLT* są bardzo zróżnicowane i wartości dość znacząco odbiegają od średnich. W przypadku pozostałych kolumn dane są wyraźnie skoncentrowane wokół średniej.

2.3 Ocena zgodności danych z rozkładem normalnym

Niektóre testy statystyczne wymagają, żeby dane miały rozkład normalny. W tym podrozdziale zostanie sprawdzony rozkład wartości kolumn dla każdej z grup. W tym celu został wykorzystany

kod z listingu 2.5. Wyniki zostały przedstawione w tabeli 2.5. Wartość „TAK” w tabeli oznacza, że dla danej grupy dane w kolumnie mają rozkład normalny, natomiast „NIE” oznacza, że nie mają.

```
1 for (i in c("KONTROLA", "CHOR1", "CHOR2")){
2     for(j in c("hsCRP", "ERY", "PLT", "HGB", "HCT", "MCHC", "MON",
        ↪ "LEU")){
3         print(i)
4         print(j);
5         print(shapiro.test(with(df, df[grupa == i, ])[[j]])$p.value
            ↪ );
6     }
7 }
```

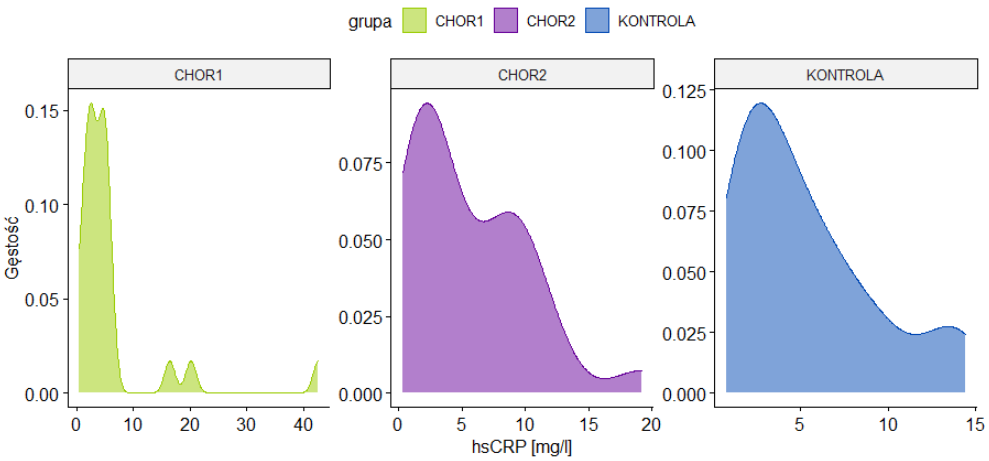
Listing 2.5: Kod sprawdzający czy w obrębie grup w poszczególnych kolumnach dane mają rozkład normalny.

	KONTROLNA	CHOR1	CHOR2
hsCRP	NIE	NIE	NIE
ERY	TAK	NIE	TAK
PLT	NIE	TAK	NIE
HGB	TAK	TAK	NIE
HCT	TAK	TAK	TAK
MCHC	TAK	TAK	TAK
MON	TAK	NIE	TAK
LEU	TAK	TAK	TAK

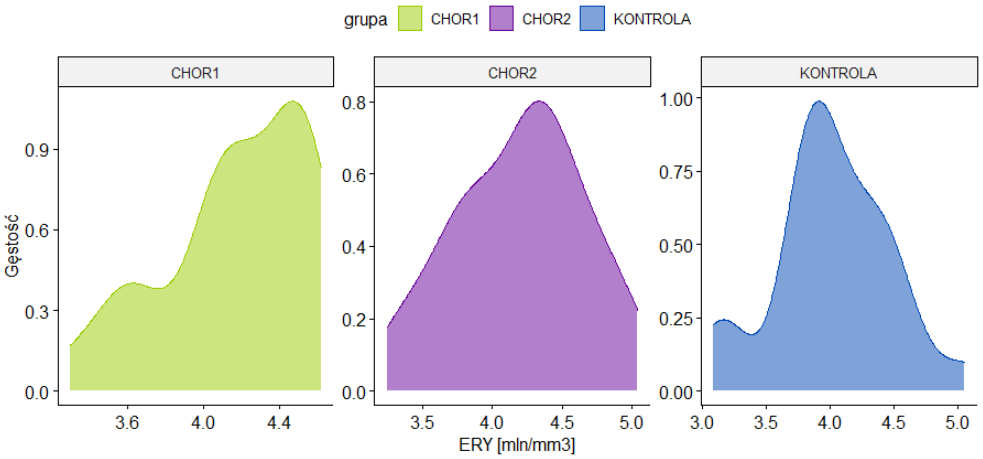
Tablica 2.5: Wyniki sprawdzenia czy dane mają rozkład normalny w obrębie grup i kolumn.

Następnie na wykresach od 2.5 do 2.9 przedstawiono rozkład danych dla tych kolumn, w których chociaż w jednej grupie nie ma rozkładu normalnego. Zaobserwowano następujące właściwości:

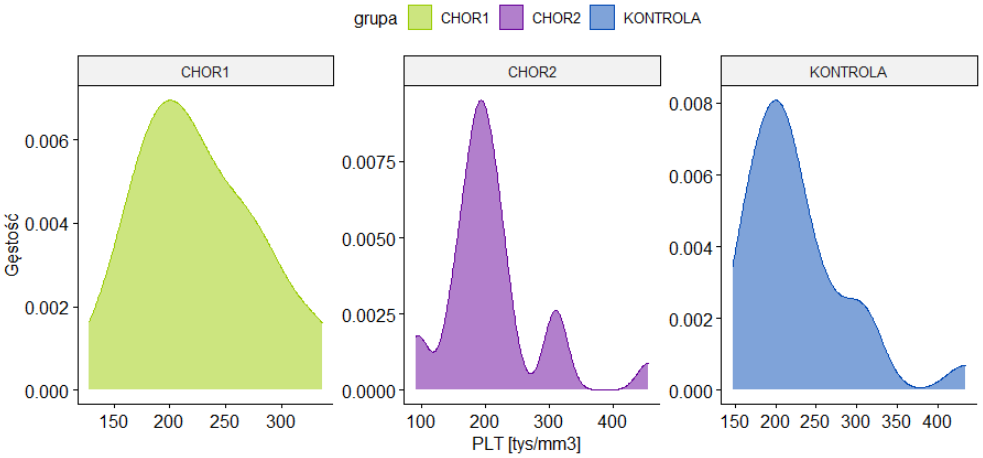
- rozkłady dla kolumny *hsCRP* dla grup *CHOR1* i *CHOR2* nie są zbliżone do normalnego, natomiast rozkład dla grupy *KONTROLA* jest zbliżony do normalnego, każdy w rozkładów jest przesunięty w kierunku mniejszych wartości,
- rozkład dla kolumny *ERY* dla grupy *CHOR1* jest przesunięty w kierunku większych wartości,
- rozkłady dla kolumny *PLT* dla grup *CHOR2* oraz *KONTROLA* są zbliżone do normalnego, ale są przesunięte w kierunku mniejszych wartości,
- rozkład dla kolumny *HGB* dla grupy *CHOR2* nie jest zbliżony do normalnego, wartości są skupione w obrębie 12 [g/dl],
- rozkład dla kolumny *MON* dla grupy *CHOR1* jest zbliżony do normalnego, ale jest przesunięty w kierunku mniejszych wartości.



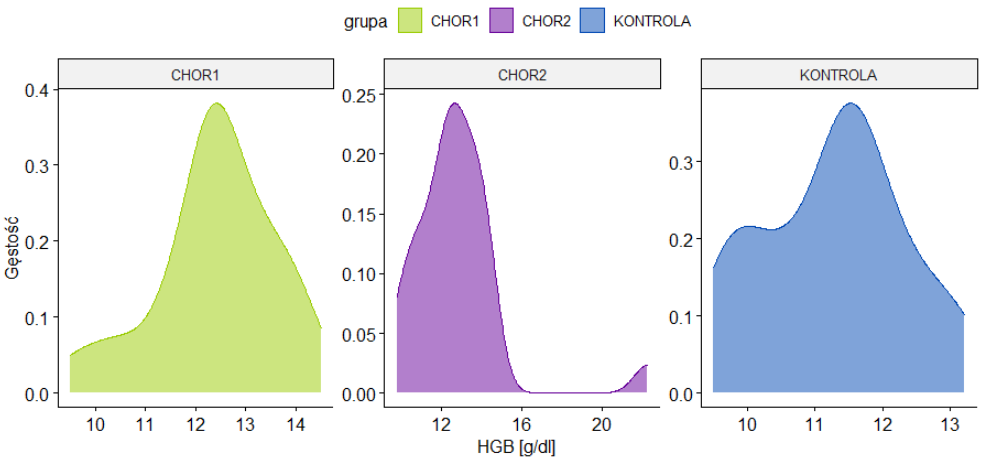
Rysunek 2.5: Wykres rozkładu dla kolumny hsCRP z podziałem na grupy.



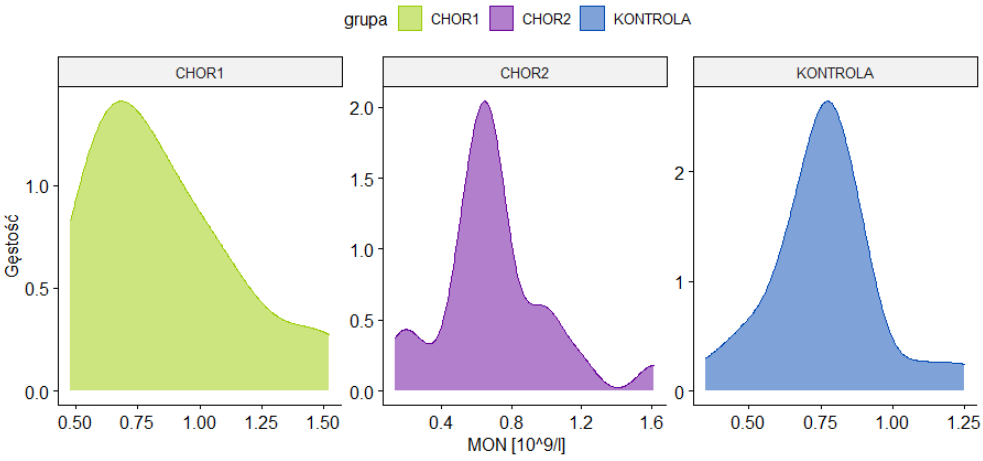
Rysunek 2.6: Wykres rozkładu dla kolumny ERY z podziałem na grupy.



Rysunek 2.7: Wykres rozkładu dla kolumny PLT z podziałem na grupy.



Rysunek 2.8: Wykres rozkładu dla kolumny HGB z podziałem na grupy.



Rysunek 2.9: Wykres rozkładu dla kolumny MON z podziałem na grupy.

2.4 Ocena homogeniczności wariancji

Ze względu na fakt, że niektóre testy statystyczne wymagają nie tylko zgodności danych z rozkładem normalnym, ale także homogeniczności wariancji. Została przeprowadzona ocena homogeniczności dla kolumn: *HCT*, *MCHC* i *LEU*. Wyniki zostały przedstawione w tabeli 2.6. We wszystkich trzech przypadkach wariancja jest homogeniczna.

	Czy wariancja homogeniczna?
HCT	TAK
MCHC	TAK
LEU	TAK

Tablica 2.6: Wyniki sprawdzenia homogeniczności wariancji dla kolumn: *HCT*, *MCHC* i *LEU*.

Rozdział 3

Analiza porównawcza między grupami

3.1 Analiza porównawcza między kolumnami mającymi rozkład normalny i spełniającymi warunek homogeniczności wariancji

Ze względu na fakt, że badane dane pochodzą z trzech grup osób oraz dane są niezależne, na tym etapie zostanie wykonany test *ANOVA* dla kolejnych kolumn: *HCT*, *MCHC* i *LEU*. Wyniki zostały przedstawione w tabeli 3.1. Wyniki wskazują, że dla kolumn *HCT* oraz *MCHC* występują różnice między grupami.

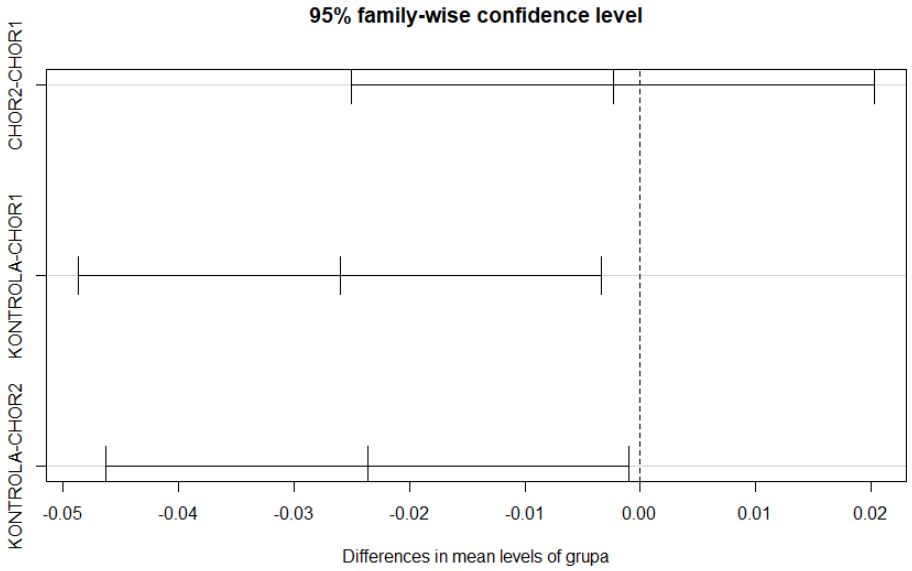
	Czy są różnice między grupami?
HCT	TAK
MCHC	TAK
LEU	NIE

Tablica 3.1: Wyniki przeprowadzenia testu ANOVA dla kolumn: *HCT*, *MCHC* i *LEU*.

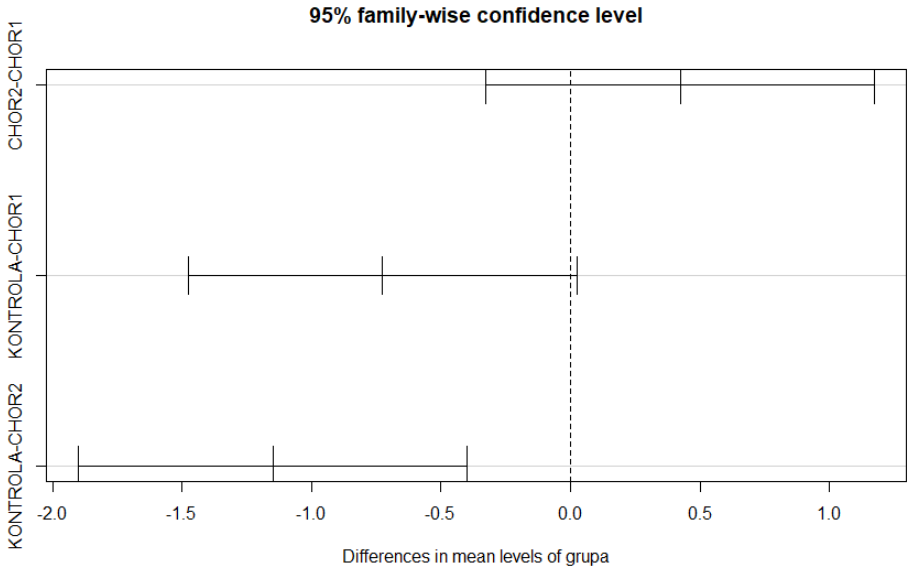
W związku z powyższymi wynikami przeprowadzono testy *post hoc TukeyHSD* dla kolumn *HCT* oraz *MCHC*, aby określić pomiędzy którymi grupami występują różnice. Wyniki zostały przedstawione na wykresie 3.1 i 3.2, natomiast wartości p w tabeli 3.2.

	HCT	**MCHC**
CHOR2-CHOR1	0.966	0.372
KONTROLA-CHOR1	0.020	0.060
KONTROLA-CHOR2	0.038	0.001

Tablica 3.2: Wartości p dla kolumn: *HCT* i *MCHC* oraz grup.



Rysunek 3.1: Wykres testu *TukeyHSD* dla kolumny *HCT*.



Rysunek 3.2: Wykres testu *TukeyHSD* dla kolumny *MCHC*.

Na podstawie wartości p oraz wykresów można stwierdzić, że:

- dla kolumny *HCT* występują istotne różnice w obrębie grup *KONTROLA - CHOR1* oraz *KONTROLA - CHOR2*,
- dla kolumny *HCT* nie występują istotne różnice w obrębie grup *CHOR2 - CHOR1*,
- dla kolumny *MCHC* występują istotne różnice w obrębie grup *KONTROLA - CHOR2*,
- dla kolumny *MCHC* nie występują istotne różnice w obrębie grup *KONTROLA - CHOR1* oraz *CHOR2 - CHOR1*.

3.2 Analiza pozostałych grup

Na tym etapie zostały zbadane dane, dla których przynajmniej jedna kolumna nie miała rozkładu normalnego: *hsCRP*, *ERY*, *PLT*, *HGB* oraz *MON*. Pierwszym przeprowadzonym testem był test *Kruskala-Wallisa*, którego wyniki zostały przedstawione w tabeli 3.3.

	Czy są różnice między grupami?
hsCRP	NIE
ERY	NIE
PLT	NIE
HGB	TAK
MON	NIE

Tablica 3.3: Wyniki przeprowadzenia testu *Kruskala-Wallisa* dla poszczególnych kolumn.

Następnie przeprowadzono test *Dunna* dla kolumny *HGB*, którego wyniki został przedstawione w tabeli 3.4.

	Wartość *p*
CHOR1-CHOR2	0.904
CHOR1-KONTROLA	0.002
CHOR2-KONTROLA	0.002

Tablica 3.4: Wyniki przeprowadzenia testu *Dunna* dla kolumny *HGB*.

Na podstawie wartości p w tabeli 3.4 można stwierdzić, że występują różnice dla parametru *HGB* w obrębie grup *CHOR1-KONTROLA* oraz *CHOR2-KONTROLA*, natomiast w obrębie grupy *CHOR1-CHOR2* nie zaobserwowano różnic.

Rozdział 4

Analiza korelacji

W ostatnim rozdziale przeprowadzono analizę korelacji najpierw dla parametrów mających rozkład normalny, w dalszej części - dla pozostałych.

4.1 Analiza korelacji dla parametrów z rozkładem normalnym

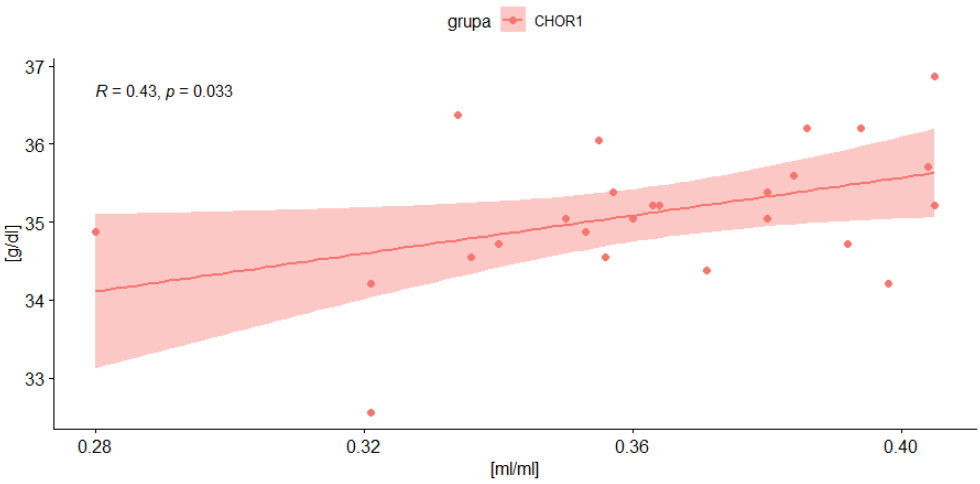
W tym podrozdziale, przy pomocy korelacji liniowej *Persona*, przeprowadzono analizę korelacji dla parametrów: *HCT*, *MCHC* i *LEU*. Wyniki zostały przedstawione w tabeli 4.1, na zielono zostały oznaczone wartości, dla których wartość p wskazywała na istotność statystyczną. Jedynymi parametrami, który wykazywały korelację były *HCT*, *MCHC* w grupie *CHOR1*. Wykres korelacji został przedstawiony na wykresie 4.1. Z wykresu wynika, że przedstawiona korelacja ma charakter dodatni, o średnim natężeniu - wraz ze wzrostem wartości parametru *HCT* wzrasta wartość parametru *MCHC*.

CHOR1			
	HCT	**MCHC**	**LEU**
HCT	-		
MCHC	0.4268	-	
LEU	-0.056	0.0836	-

CHOR2			
	HCT	**MCHC**	**LEU**
HCT	-		
MCHC	-0.118	-	
LEU	0.3072	-0.012	-

KONTROLNA			
	HCT	**MCHC**	**LEU**
HCT	-		
MCHC	-0.048	-	
LEU	-0.360	0.1071	-

Tablica 4.1: Wyniki analizy korelacji dla parametrów: *HCT*, *MCHC* i *LEU*.



Rysunek 4.1: Wykres korelacji parametrów *HCT* i *MCHC* w grupie *CHOR1*.

4.2 Analiza korelacji dla pozostałych parametrów

W tym podrozdziale, przy pomocy korelacji rang *Spearmana*, przeprowadzono analizę korelacji dla parametrów: *hsCRP*, *ERY*, *PLT*, *HGB* i *MON*. Wyniki zostały przedstawione w tabeli 4.2, na zielono zostały oznaczone wartości, dla których wartość p wskazywała na istotność statystyczną.

CHOR1					
	hsCRP	**ERY**	**PLT**	**HGB**	**MON**
hsCRP	-				
ERY	0.0365	-			
PLT	-0.058	-0.066	-		
HGB	-0.121	0.875	-0.181	-	
MON	0.3818	-0.222	-0.053	-0.265	-

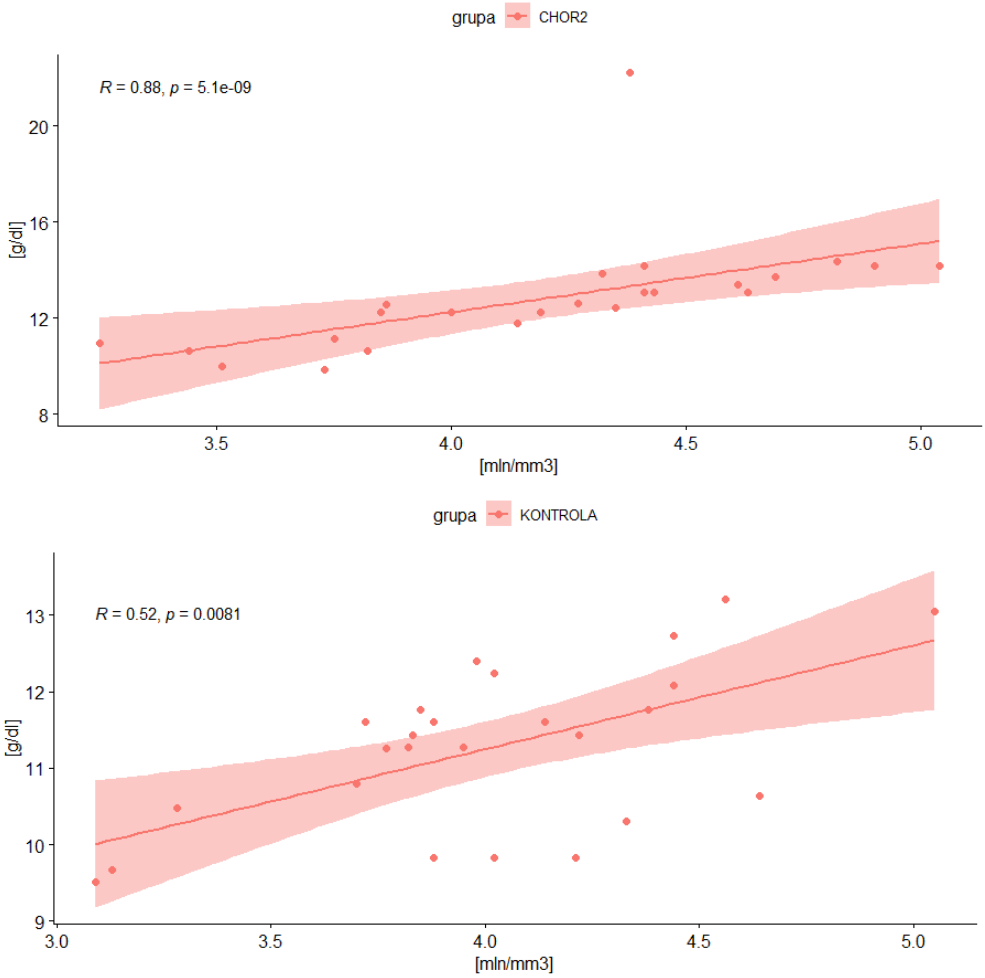
CHOR2					
	hsCRP	**ERY**	**PLT**	**HGB**	**MON**
hsCRP	-				
ERY	0.1004	-			
PLT	0.2340	0.0800	-		
HGB	-0.032	0.8831	-0.060	-	
MON	0.0770	-0.310	0.0157	-0.387	-

KONTROLNA					
	hsCRP	**ERY**	**PLT**	**HGB**	**MON**
hsCRP	-				
ERY	0.2890	-			
PLT	0.1315	0.4542	-		
HGB	0.1896	0.5175	0.3155	-	
MON	-0.005	-0.132	0.1478	-0.275	-

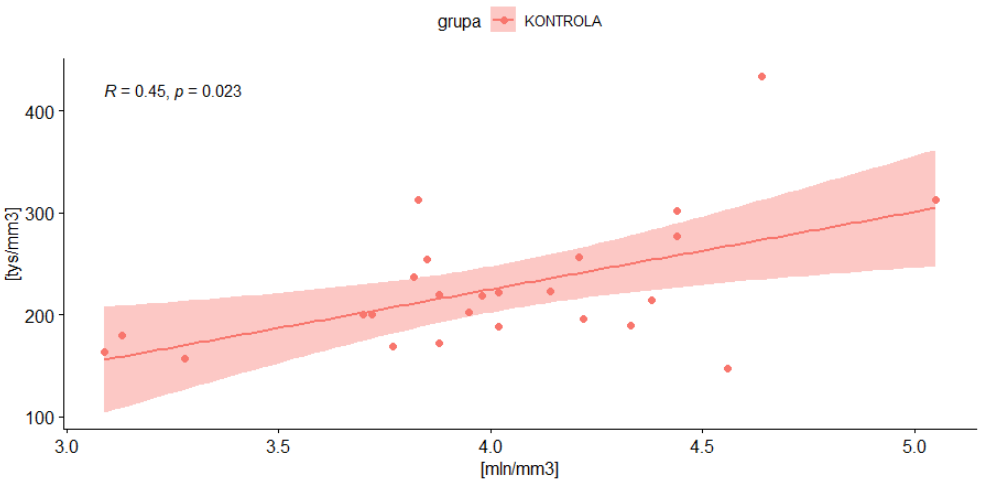
Tablica 4.2: Wyniki analizy korelacji dla pozostałych parametrów.

We wszystkich trzech grupach występowała silna lub bardzo silna korelacja między warto-

ścią parametru *ERY*, a wartością parametru *HGB*. Porównanie korelacji tych dwóch parametrów między grupami *CHOR2* i *KONTROLNA* zostało przedstawione na wykresie 4.2. Dodatkowo w grupie *KONTROLNA* wystąpiła korelacja dodatnia o średnim natężeniu między parametrami *ERY* i *PLT*, co zostało przedstawione na wykresie 4.3.



Rysunek 4.2: Porównanie wykresów korelacji parametrów *ERY* i *HGB* między grupami *CHOR2* i *KONTROLA*.



Rysunek 4.3: Wykres korelacji parametrów *ERY* i *PLT* w grupie *KONTROLA*.

Rozdział 5

Podsumowanie

W ramach projektu opracowano dane dotyczące badania krwi, przeprowadzono ich statystykę opisową, analizę porównawczą oraz analizę korelacji. Na kolejnych etapach przedstawiono zaobserwowane właściwości danych. Brak konkretnych wniosków wynika z braku wiedzy na temat eksperymentu oraz wiedzy dziedzinowej, jednak przygotowane sprawozdanie powinno pozwolić osobie z takową wiedzą na postawienie konkretnych wniosków lub zlecenie badań pozwalających na zbadanie zaobserwowanych własności statystycznych.

Literatura

[1] Przykładowe dane. [on-line] www.cs.put.poznan.pl/kgutowska/PSwBB/dane/przykladoweDane-Projekt.csv.

[2] Hematokryt. [on-line] https://lekarzebezkolejki.pl/blog/hematokryt-hct-normy-co-oznacza-niski-i-wysoki-poziom-hct/w-603.

[3] Białko hscrp. [on-line] https://www.labtestsonline.pl/test/hs-crp.

[4] Mchc. [on-line] https://www.medonet.pl/zdrowie/pytania-do-lekarzy,o-czym-swiadczy-mchc-ponizej-normy-,porada,43426672.html.

[5] Dopuszczalne zakresy różnych składników krwi. [on-line] https://www.medonet.pl/.