Sprawozdanie - Statystyczna Analiza Danych

Julia Ordecka

Bioinformatyka, II rok, grupa I

Skrypt przeprowadza analizę statystyczną dla danych niezależnych.

1. Przygotowanie danych wejściowych

W przypadku braków w danych wejściowych brakujące wartości (NA) są zastępowane za pomocą funkcji impute. W każdej w ramce danych z danymi numerycznymi brakujące dane zostają zastąpione wartością średnią z danej kolumny dla odpowiadającej grupy.
Kod grupuje dane według unikalnych wartości w kolumnie określonej przez group_col i stosuje imputację wartości średniej do wszystkich kolumn numerycznych.

Fragment ramki danych przed zastąpieniem braków:

	grupa	plec	wiek	hsCRP	ERY	PLT	HGB	HCT	MCHC	MON	LEU
1	CHOR1	k	36	2.711000	4.19	201	13.21020	0.3920	34.71490	0.48	11.86
2	CHOR1	m	39	4.699380	4.48	222	13.04910	0.3800	35.37930	0.76	10.32
3	CHOR1	k	35	2.353540	3.59	278	10.14930	0.3210	32.55560	1.08	13.60
4	CHOR1	m	29	2.271610	3.66	200	11.27700	0.3360	34.54880	0.63	10.11
5	CHOR1	m	29	4.465190	4.41	128	12.40470	0.3630	35.21320	*NA*	10.55
6	CHOR1	m	43	6.162690	3.68	176	11.43810	0.3400	34.71490	0.83	9.28
7	CHOR1	k	29	4.988360	4.12	288	12.24360	0.3570	35.37930	0.90	10.07
8	CHOR1	k	26	1.849380	4.44	231	13.21020	0.3980	34.21660	0.74	9.56
9	CHOR1	m	23	20.154800	4.13	153	12.56580	0.3840	35.59523	1.07	14.48
10	CHOR1	m	23	3.204050	4.02	249	11.92140	0.3530	34.88100	1.07	10.51
11	CHOR1	m	24	0.487607	4.07	177	11.92140	0.3500	35.04710	0.61	6.79
12	CHOR1	k	30	2.322680	4.11	295	12.24360	0.3600	35.04710	0.72	14.97
13	CHOR1	m	26	16.406900	4.18	174	*NA*	0.3340	36.37590	1.50	16.00
14	CHOR1	k	27	3.044270	4.59	207	13.85460	0.3940	36.20980	0.59	9.23

Fragment ramki danych po imputowaniu:

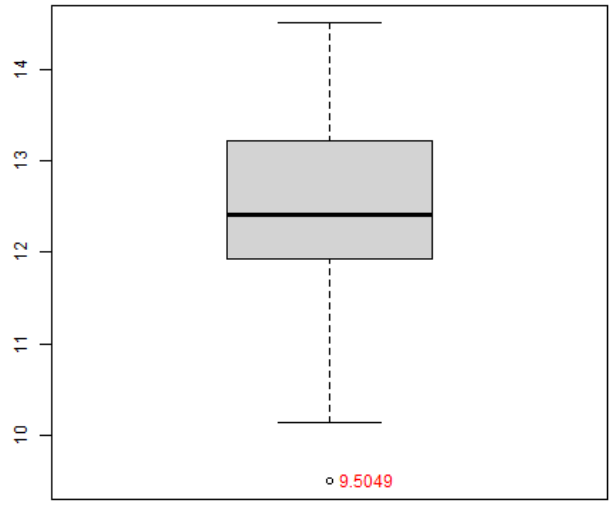
	grupa	plec	wiek	hsCRP	ERY	PLT	HGB	HCT	MCHC	MON	LEU
1	CHOR1	k	36	2.711000	4.19	201	13.21020	0.3920	34.71490	0.4800000	11.86
2	CHOR1	m	39	4.699380	4.48	222	13.04910	0.3800	35.37930	0.7600000	10.32
3	CHOR1	k	35	2.353540	3.59	278	10.14930	0.3210	32.55560	1.0800000	13.60
4	CHOR1	m	29	2.271610	3.66	200	11.27700	0.3360	34.54880	0.6300000	10.11
5	CHOR1	m	29	4.465190	4.41	128	12.40470	0.3630	35.21320	0.8579167	10.55
6	CHOR1	m	43	6.162690	3.68	176	11.43810	0.3400	34.71490	0.8300000	9.28
7	CHOR1	k	29	4.988360	4.12	288	12.24360	0.3570	35.37930	0.9000000	10.07
8	CHOR1	k	26	1.849380	4.44	231	13.21020	0.3980	34.21660	0.7400000	9.56
9	CHOR1	m	23	20.154800	4.13	153	12.56580	0.3840	35.59523	1.0700000	14.48
10	CHOR1	m	23	3.204050	4.02	249	11.92140	0.3530	34.88100	1.0700000	10.51
11	CHOR1	m	24	0.487607	4.07	177	11.92140	0.3500	35.04710	0.6100000	6.79
12	CHOR1	k	30	2.322680	4.11	295	12.24360	0.3600	35.04710	0.7200000	14.97
13	CHOR1	m	26	16.406900	4.18	174	12.41141	0.3340	36.37590	1.5000000	16.00
14	CHOR1	k	27	3.044270	4.59	207	13.85460	0.3940	36.20980	0.5900000	9.23

Przygotowano raport wartości odstających dla danych parametrów:

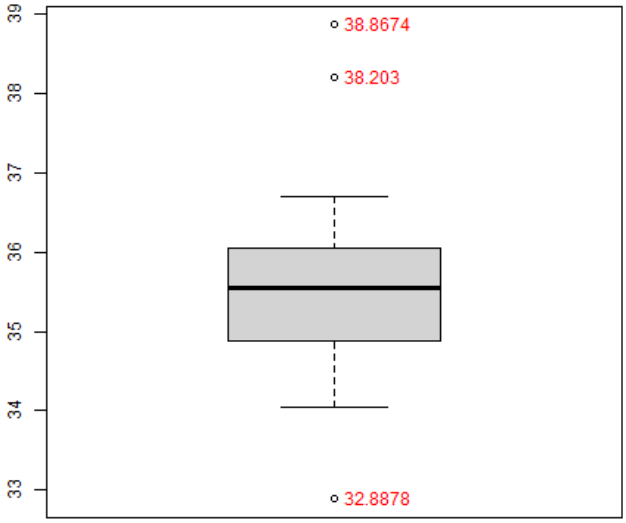
```
[1] "Raport wartości odstających:"
   parameter outlier_value
1       wiek       48.0000
2      hsCRP       20.1548
3      hsCRP       16.4069
4      hsCRP       42.6499
5      hsCRP       19.2124
6        ERY       33.0000
7        PLT      456.0000
8        PLT      434.0000
9        HGB       22.2318
10       HCT        0.0423
11      MCHC       38.8674
12      MCHC       38.2030
13      MCHC       32.0573
14      MCHC       32.2234
15       MON        1.5000
16       MON        1.5200
17       MON        0.1400
18       MON        1.6100
19       MON        7.0000
```

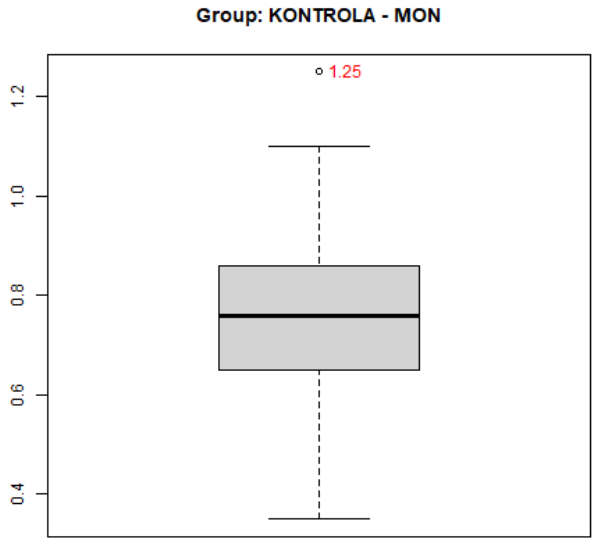
Poniżej przedstawiono przykładowe wykresy wizualizujące wartości odstające dla wybranych parametrów za pomocą boxplotów z wyznaczeniem wartości odstających w poszczególnych grupach.

Group: CHOR1 - HGB



Group: CHOR2 - MCHC





Na wykresach boxplot wartości odstające zostały oznaczone okręgami oraz podpisane.

2. **Wykonanie charakterystyk dla badanych grup**

   Charakterystyki badanych grup dokonuje się poprzez wykorzystanie statystyk opisowych w ocenie parametrów.
   Za pomocą funkcji generate_summary skrypt generuje statystyki ogólne dla wszystkich kolumn w pliku wejściowym. Statystyki ogólne zawierają takie informacje jak częstość występowania danej wartości dla danych nienumerycznych oraz poniżej wymienione metryki dla danych numerycznych:
   Min: minimalna wartość w kolumnie,
   1st Qu: pierwszy kwartyl (25. percentyl) w kolumnie,
   Median: mediana w kolumnie,
   Mean: Średnia arytmetyczna w kolumnie,
   3rd Qu: Trzeci kwartyl w kolumnie,
   Max: Maksymalna wartość w kolumnie.

   Funkcja podsumowanie_kolumn tworzy statystyki grupowe - podsumowanie dla każdej kolumny numerycznej w ramce danych według danej kolumny grupującej group_col i zapisuje te podsumowania jako osobne ramki danych. Wyniki są przechowywane w osobnej ramce danych dla każdej kolumny.

   Statystyki grupowe zawierają informacje o następujących wartościach:

   count: liczba obserwacji (wierszy) w każdej grupie
   min: Minimalna wartość w kolumnie
   median: Mediana

mean: Średnia arytmetyczna
max: Maksymalna wartość
sd: Odchylenie standardowe
IQR: Rozstęp międzykwartylowy (IQR)
var: Wariancja

Przykładowe ramki danych z podsumowaniem ogólnym:

	Value	Frequency	Variable
1	k	40	plec
2	m	35	plec

	Var1	Var2	Freq	Variable
1	A	Min.	9.50490	HGB
2	A	1st Qu.	11.27700	HGB
3	A	Median	12.24360	HGB
4	A	Mean	12.16923	HGB
5	A	3rd Qu.	13.04910	HGB
6	A	Max.	22.23180	HGB

	Var1	Var2	Freq	Variable
1	A	Min.	32.05730	MCHC
2	A	1st Qu.	34.38270	MCHC
3	A	Median	35.04710	MCHC
4	A	Mean	35.02783	MCHC
5	A	3rd Qu.	35.71150	MCHC
6	A	Max.	38.86740	MCHC

Podsumowanie dla wybranych danych numerycznych względem grup:

	grupa	variable	count	min	median	mean	max	sd	IQR	var
1	CHOR1	MCHC	25	32.5556	35.0471	35.12882	36.8742	0.8775039	0.88033	0.770013
2	CHOR2	MCHC	25	32.8878	35.5454	35.55204	38.8674	1.2906016	1.16270	1.665653
3	KONTROLA	MCHC	25	32.0573	34.5488	34.40263	36.0437	1.1197078	1.49490	1.253746

	grupa	variable	count	min	median	mean	max	sd	IQR	var
1	CHOR1	hsCRP	25	0.487607	3.96646	6.103022	42.6499	8.824633	2.67086	77.87415
2	CHOR2	hsCRP	25	0.335089	3.44546	5.536029	19.2124	4.645587	6.53121	21.58148
3	KONTROLA	hsCRP	25	0.758440	4.22037	5.295149	14.3951	3.996580	4.54994	15.97265

	grupa	variable	count	min	median	mean	max	sd	IQR	var
1	CHOR1	MON	25	0.48	0.76	0.8579167	1.52	0.2992451	0.46	0.08954764
2	CHOR2	MON	25	0.14	0.66	0.9528000	7.00	1.2978685	0.33	1.68446267
3	KONTROLA	MON	25	0.35	0.76	0.7604000	1.25	0.1875162	0.21	0.03516233

3. Analiza porównawcza pomiędzy grupami

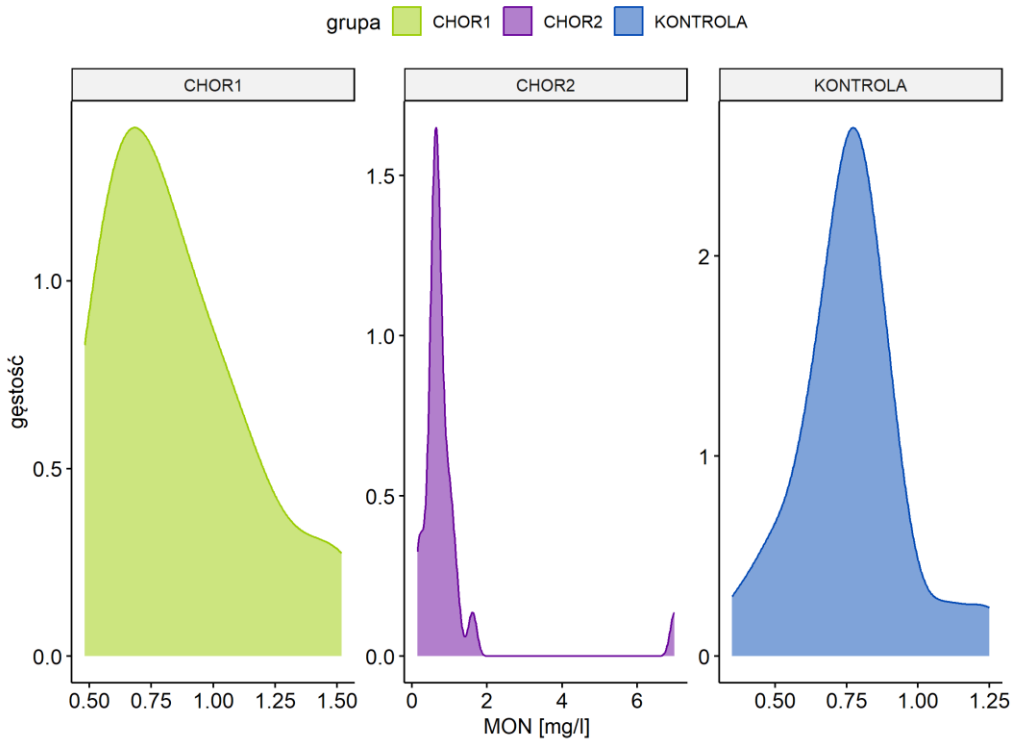
Porównywane są grupy niezależne - badane są te same parametry u różnych grup – kontrolnych i grup chorych.
Grupy są o liczebności >2.

Do zbadania zgodności z rozkładem normalnym wykorzystano test Shapiro-Wilka. Wartość p-value większa od 0,05 oznacza, że dane są zgodne z rozkładem normalnym. Wartość p-value mniejsza od 0,05 świadczy o niezgodności danych z rozkładem normalnym. Również wartość statystyki testowej wskazuje na zgodność z rozkładem normalnym - wartości bliskie 1 wskazują, że rozkład danych jest bardziej zbliżony do rozkładu normalnego.

Fragment ramki danych zawierający wyniki testu Shapiro-Wilka:

grupa	variable	statistic	p_value	normality
CHOR1	wiek	0.9691404	6.233499e-01	zgodny z rozkładem normalnym
CHOR2	wiek	0.9569757	3.575393e-01	zgodny z rozkładem normalnym
KONTROLA	wiek	0.9551495	3.263464e-01	zgodny z rozkładem normalnym
CHOR1	hsCRP	0.5418468	9.933661e-08	niezgodny
CHOR2	hsCRP	0.8731146	4.998515e-03	niezgodny
KONTROLA	hsCRP	0.8813002	7.351605e-03	niezgodny
CHOR1	ERY	0.2494035	2.653537e-10	niezgodny
CHOR2	ERY	0.9815472	9.139531e-01	zgodny z rozkładem normalnym
KONTROLA	ERY	0.9692766	6.267790e-01	zgodny z rozkładem normalnym
CHOR1	PLT	0.9662912	5.532056e-01	zgodny z rozkładem normalnym
CHOR2	PLT	0.8636689	3.237412e-03	niezgodny
KONTROLA	PLT	0.8675775	3.869518e-03	niezgodny
CHOR1	HGB	0.9643358	5.073768e-01	zgodny z rozkładem normalnym
CHOR2	HGB	0.7513306	3.885087e-05	niezgodny
KONTROLA	HGB	0.9589733	3.944054e-01	zgodny z rozkładem normalnym

Wygenerowano wykresy gęstości, które można zinterpretować w kontekście zgodności z rozkładem normalnym. Jeśli wykres gęstości jest symetryczny względem środka (średniej) i ma jeden szczyt, to może to sugerować, że rozkład danych jest zbliżony do rozkładu normalnego.
Przykładowy wykres dla parametru MON:



Z wykresów wynika, że dla parametru MON w grupie CHOR1 i CHOR2 dane są niezgodne z rozkładem normalnym.

Do analizy homogeniczności wariancji wykorzystano test Levene'a. Dla wartości p-value wynoszącej >0,05 można założyć jednorodność wariancji.

Zgodnie z poniższą tabelą dobrane zostały testy statystyczne:

Porównanie grup niezależnych			
Ilość porównywanych grup	Zgodność z rozkładem normalnym	Jednorodność wariancji	Wybrany test
2	TAK	TAK	test t-Studenta (dla gr. niezależnych)
		NIE	test Welcha
	NIE	-	test Wilcoxona (Manna-Whitneya)
>2	TAK	TAK	test ANOVA (*post hoc* Tukeya)
		NIE	test Kruskala-Wallisa (*post hoc* Dunna)
	NIE	-	

Jeśli dane nie spełniają założenia o zgodności z rozkładem normalnym (p-value < 0.05) do analizy porównawczej wykorzystuje się testy nieparametryczne, np. test Kruskala-Wallisa. Tak samo w przypadku, gdy dane są zgodne z rozkładem normalnym, ale nie spełniają założenia o jednorodności wariancji to również stosuje się test Kruskala-Wallisa.

W przypadku niespełnienia założenia dotyczącego zgodności z rozkładem normalnym lub w przypadku spełnienia tego założenia, ale jednocześnie niespełnienia założenia o jednorodności wariancji wykorzystuje się test Kruskala-Wallisa. W przypadku spełnienia warunku zgodności z rozkładem normalnym oraz warunku jednorodności wariancji stosuje się test ANOVA.

variable	test_type	p_value
wiek	ANOVA	0.2056278453
hsCRP	Kruskal-Wallis	0.8807212193
ERY	Kruskal-Wallis	0.1543513635
PLT	Kruskal-Wallis	0.3240306971
HGB	Kruskal-Wallis	0.0007673315
HCT	Kruskal-Wallis	0.0189609134
MCHC	ANOVA	0.0018598103
MON	Kruskal-Wallis	0.2542134912
LEU	ANOVA	0.5965009414

Gdy wynikowa wartość p-value wynosi <0,05 można założyć, że istnieją znaczące różnice pomiędzy badanymi grupami. Stosując testy post hoc można ocenić pomiędzy którymi grupami są istotne różnice i jak one są duże.

Przykładowy wynik testu post hoc dla parametru MCHC:

	diff	lwr	upr	p adj	Comparison	Variable
CHOR2-CHOR1	0.4232228	-0.3274109	1.17385653	0.372940393	CHOR2-CHOR1	MCHC
KONTROLA-CHOR1	-0.7261892	-1.4768229	0.02444453	0.060043259	KONTROLA-CHOR1	MCHC
KONTROLA-CHOR2	-1.1494120	-1.9000457	-0.39877827	0.001352322	KONTROLA-CHOR2	MCHC

-diff to różnica średnich między dwiema porównywanymi grupami
-lwr to dolna granica przedziału ufności dla różnicy średnich
-upr to górna granica przedziału ufności dla różnicy średnich
-p adj mówi, czy różnica między grupami jest statystycznie istotna

CHOR2-CHOR1: różnica nie jest statystycznie istotna przy poziomie istotności 0.05 (p adj = 0.373)

KONTROLA-CHOR1: różnica ta jest bliska istotności, ale nie jest statystycznie istotna

KONTROLA-CHOR2: różnica jest statystycznie istotna

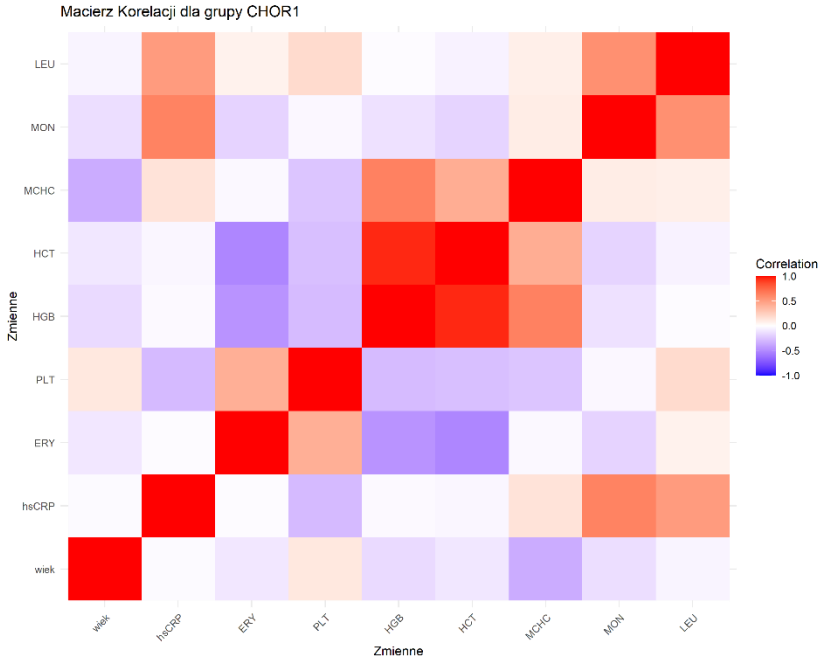
4. Analiza korelacji

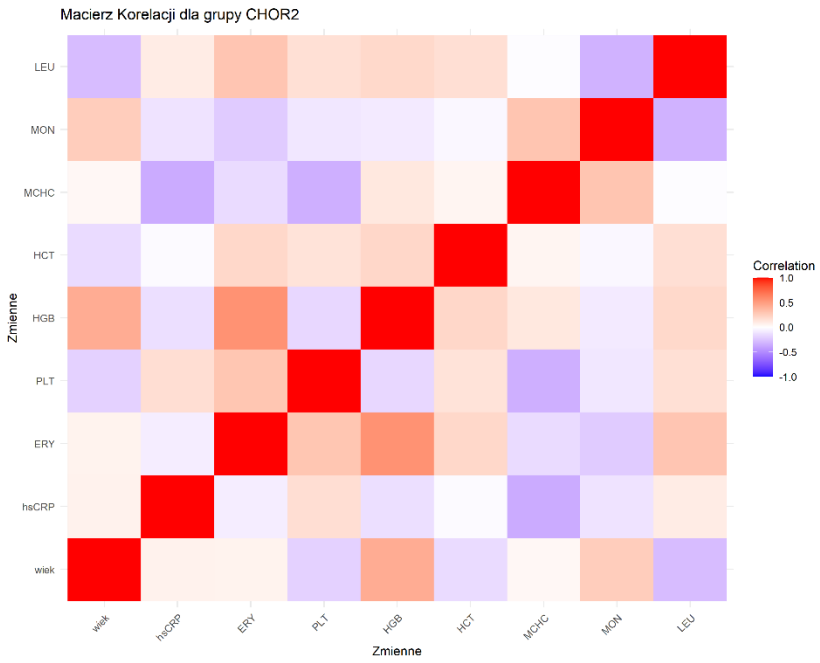
- $-1 < r \leq -0.7$ bardzo silna korelacja ujemna
- $-0.7 < r \leq -0.5$ silna korelacja ujemna
- $-0.5 < r \leq -0.3$ korelacja ujemna o średnim natężeniu
- $-0.3 < r \leq -0.2$ słaba korelacja ujemna
- $-0.2 < r < 0.2$ brak korelacji
- $0.2 \leq r < 0.3$ słaba korelacja dodatnia
- $0.3 \leq r < 0.5$ korelacja dodatnia o średnim natężeniu
- $0.5 \leq r < 0.7$ silna korelacja dodatnia
- $0.7 \leq r < 1$ bardzo silna korelacja dodatnia

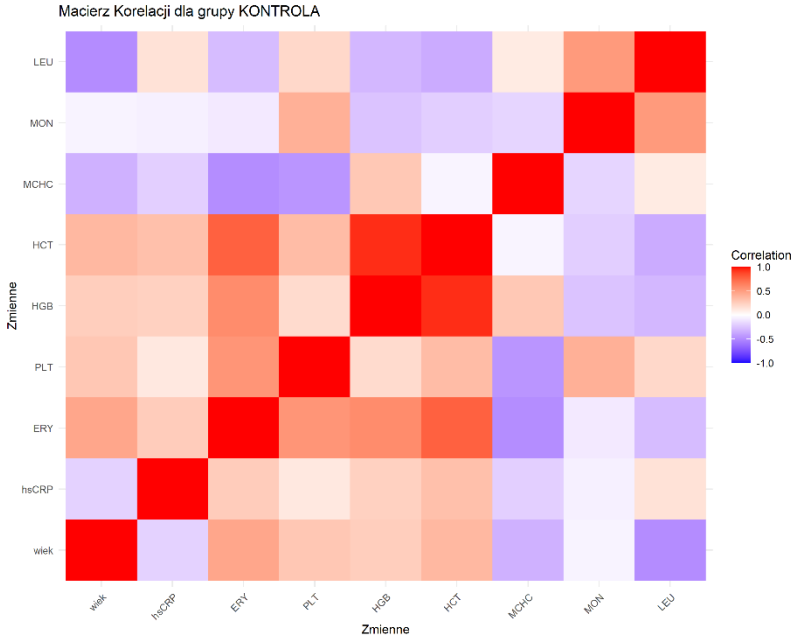
Dla danych parametrycznych użyty został współczynnik korelacji liniowej Pearsona. W przypadku danych nieparametrycznych użyto współczynnika korelacji rangowej Spearmana.
Fragment ramki danych:

Group	parametr1	parametr2	Test	Est	p_value	interpretacja
CHOR1	hsCRP	MCHC	spearman	0.1457776234	4.868672e-01	brak korelacji
CHOR1	hsCRP	MON	spearman	0.3818322917	5.963262e-02	korelacja dodatnia o średnim natężeniu
CHOR1	hsCRP	LEU	spearman	0.2676923077	1.950876e-01	słaba korelacja dodatnia
CHOR1	ERY	PLT	spearman	0.1454405648	4.878860e-01	brak korelacji
CHOR1	ERY	HGB	spearman	0.6443040352	5.088237e-04	silna korelacja dodatnia
CHOR1	ERY	HCT	spearman	0.6150144484	1.068964e-03	silna korelacja dodatnia
CHOR1	ERY	MCHC	spearman	0.3699855592	6.869226e-02	korelacja dodatnia o średnim natężeniu
CHOR1	ERY	MON	spearman	-0.3479014505	8.835189e-02	korelacja ujemna o średnim natężeniu
CHOR1	ERY	LEU	spearman	0.0665640680	7.519023e-01	brak korelacji
CHOR1	PLT	HGB	spearman	-0.1815382765	3.851371e-01	brak korelacji
CHOR1	PLT	HCT	spearman	-0.1200692907	5.675349e-01	brak korelacji
CHOR1	PLT	MCHC	spearman	-0.2298504326	2.690315e-01	słaba korelacja ujemna
CHOR1	PLT	MON	spearman	-0.0531177986	8.009100e-01	brak korelacji
CHOR1	PLT	LEU	spearman	0.1176923077	5.738419e-01	brak korelacji
CHOR1	HGB	HCT	spearman	0.9361754294	6.306757e-12	bardzo silna korelacja dodatnia
CHOR1	HGB	MCHC	spearman	0.5835749773	2.196412e-03	silna korelacja dodatnia
CHOR1	HGB	MON	spearman	-0.2653811483	1.998043e-01	słaba korelacja ujemna
CHOR1	HGB	LEU	spearman	0.0219696003	9.169817e-01	brak korelacji
CHOR1	HCT	MCHC	spearman	0.4175197518	3.783346e-02	korelacja dodatnia o średnim natężeniu
CHOR1	HCT	MON	spearman	-0.2724821926	1.875841e-01	słaba korelacja ujemna

Dane dotyczące korelacji zostają również wyeksportowane do pliku tekstowego. Dla każdej grup zostały wygenerowane macierze korelacji:




Macierz Korelacji dla grupy CHOR2
LEU
MON
MCHC
HCT
HGB
PLT
ERY
hsCRP
wiek
Zmienne
wiek
hsCRP
ERY
PLT
HGB
HCT
MCHC
MON
LEU
Zmienne
Correlation
1.0
0.5
0.0
-0.5
-1.0


Macierz Korelacji dla grupy KONTROLA
LEU
MON
MCHC
HCT
HGB
PLT
ERY
hsCRP
wiek
Zmienne
wiek
hsCRP
ERY
PLT
HGB
HCT
MCHC
MON
LEU
Zmienne
Correlation
1.0
0.5
0.0
-0.5
-1.0