Sprawozdanie z projektu

Statystyczna Analiza Danych

Przygotowała Julia Lemańska

Narzędzie do przeprowadzania podstawowej statystycznej analizy danych medycznych

Narzędzie w formie skryptu w języku R jest podzielone na cztery segmenty służące do poszczególnych etapów analizy statystycznej, gdzie badanych grup jest więcej niż dwie oraz grupy są niezależne:

1) Weryfikacja danych wejściowych – obsługa brakujących danych i wartości odstających.
2) Kalkulacja statystyk opisowych, rozkładu oraz wariancji danych.
3) Analiza porównawcza pomiędzy grupami.
4) Analiza korelacji pomiędzy parametrami w obrębie każdej grupy.

Poniżej znajdują się wyjaśnienia działania każdego z etapów analizy oraz ich danych wyjściowych.

1) Weryfikacja danych wejściowych – obsługa brakujących danych i wartości odstających.

Dane wejściowe są wczytywane z pliku w formacie .csv . Następnie obsługa danych brakujących jest wykonywana tylko w przypadku parametrów z danymi numerycznymi. Wartości brakujące są wypełniane średnią dla swojej grupy w danym parametrze, tzn. gdy znajdziemy wartość NA w kolumnie *ERY* dla grupy *KONTROLA*, wypełniamy tą komórkę średnią kolumny *ERY* dla grupy *KONTROLA*.

Wykorzystując dane statystyczne obliczone z funkcji boxplot, sprawdzone zostają wartości odstające od reszty danych dla danego parametru i grupy. Wartości te, nie są ani usuwane ani zastępowane, tylko wyłącznie zaraportowane.

Informacja o zastąpionych wartościach NA oraz wartościach odstających zostaje wyświetlona w konsoli.

```
Grupa CHOR1 w kolumnie hsCRP posiada wartości odstające: 42.6499
Grupa CHOR1 w kolumnie ERY posiada wartości odstające: 33
Grupa CHOR1 posiada wartości puste w kolumnie HGB - zastąpiono średnią dla tej grupy.
Grupa CHOR1 w kolumnie HGB posiada wartości odstające: 9.5049
Grupa CHOR1 w kolumnie HCT posiada wartości odstające: 0.28
Grupa CHOR1 w kolumnie MCHC posiada wartości odstające: 32.5556
Grupa CHOR1 posiada wartości puste w kolumnie MON - zastąpiono średnią dla tej grupy.
Grupa CHOR2 w kolumnie hsCRP posiada wartości odstające: 19.2124
Grupa CHOR2 w kolumnie PLT posiada wartości odstające: 456
Grupa CHOR2 w kolumnie PLT posiada wartości odstające: 314
```

2) Kalkulacja statystyk opisowych, rozkładu oraz wariancji danych.

Dla każdej grupy i każdego parametru numerycznego wykonana jest statystyka opisowa, w której znajdują się takie statystyki opisowe jak: wartości minimalna i maksymalna w próbie, średnia, wariancja, odchylenie standardowe, mediana. Aby wyświetlić zestawienie wszystkich obliczonych statystyk dla danych, należy wejść w tabelę (ramkę danych) o nazwie *ramka_charakterystyk* (zostanie ona wyświetlona automatycznie pod koniec tego segmentu).

	grupa	kolumna	min	max	mean	sd	var	median
1	KONTROLA	wiek	23.00	48.00	32.32	5.61	31.48	32.00
2	CHOR1	wiek	17.00	43.00	29.56	5.88	34.59	29.00
3	CHOR2	wiek	21.00	42.00	30.04	5.90	34.79	30.00
4	KONTROLA	hsCRP	0.76	14.40	5.30	4.00	15.97	4.22
5	CHOR1	hsCRP	0.49	42.65	6.10	8.82	77.87	3.97
6	CHOR2	hsCRP	0.34	19.21	5.54	4.65	21.58	3.45
7	KONTROLA	ERY	3.09	5.05	4.01	0.46	0.21	3.98
8	CHOR1	ERY	3.53	33.00	5.36	5.77	33.26	4.20
9	CHOR2	ERY	3.25	5.04	4.20	0.47	0.22	4.27
10	KONTROLA	PLT	147.00	434.00	225.88	63.81	4072.11	214.00

Dodatkowo dla każdego parametru tworzone są wykresy – boxploty z podziałem na grupy. Wraz z wykresami gęstości rozkładu, są one zapisywane do pliku *Wykresy-charakterystyka.pdf* w katalogu roboczym.

Oprócz statystyk opisowych, obliczany jest rozkład oraz wariancja dla poszczególnego parametru numerycznego z podziałem na grupy badawcze. Do sprawdzenia rozkładu użyty jest test Shapiro-Wilka, który określa, że można założyć rozkład normalny w przypadku gdy p.value > 0.05. Aby wyświetlić zestawienie wszystkich obliczonych rozkładów, należy wejść w tabelę (ramkę danych) o nazwie *rozklady* (zostanie ona wyświetlona automatycznie pod koniec tego segmentu).

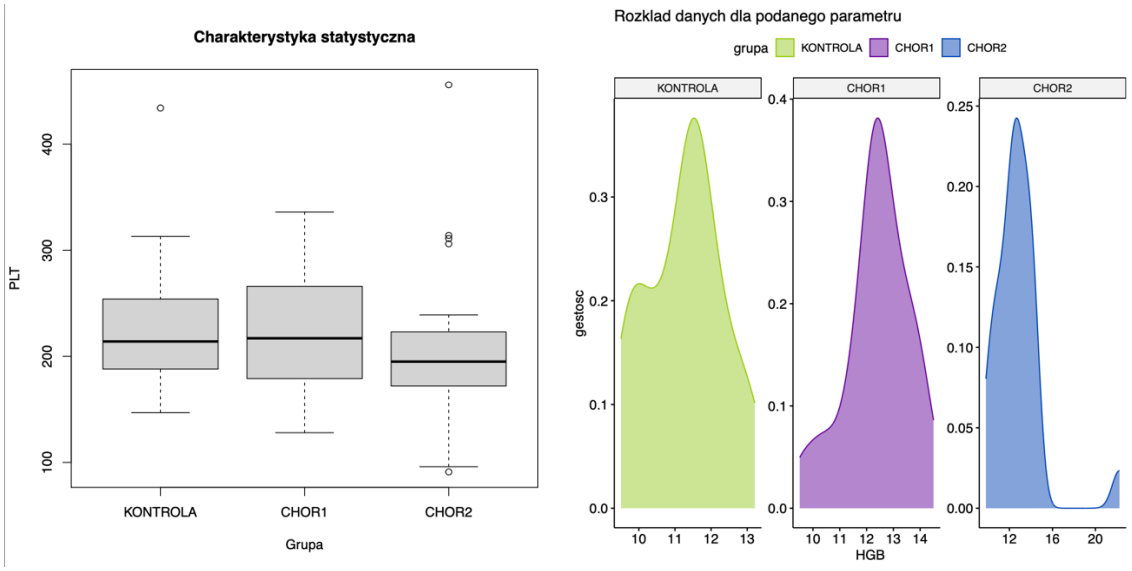
	kolumna	grupa	statistic	p.value
W...1	wiek	CHOR1	0.9691404	6.233499e-01
W...2	wiek	CHOR2	0.9569757	3.575393e-01
W...3	wiek	KONTROLA	0.9551495	3.263464e-01
W...4	hsCRP	CHOR1	0.5418468	9.933661e-08
W...5	hsCRP	CHOR2	0.8731146	4.998515e-03
W...6	hsCRP	KONTROLA	0.8813002	7.351605e-03
W...7	ERY	CHOR1	0.2494035	2.653537e-10
W...8	ERY	CHOR2	0.9815472	9.139531e-01
W...9	ERY	KONTROLA	0.9692766	6.267790e-01
W...10	PLT	CHOR1	0.9662912	5.532056e-01

Do obliczenia wariancji w danym parametrze numerycznym pomiędzy grupami wykorzystywany jest test Levene'a, który określa, że można założyć homogeniczność danych w przypadku gdy p.value > 0.05. Aby wyświetlić zestawienie wszystkich obliczonych

wariancji, należy wejść w tabelę (ramkę danych) o nazwie *wariancje* (zostanie ona wyświetlona automatycznie pod koniec tego segmentu).

	kolumna	p.value
1	wiek	0.7330851
2	hsCRP	0.7907312
3	ERY	0.4303854
4	PLT	0.9375149
5	HGB	0.1593675
6	HCT	0.1237021
7	MCHC	0.2688996
8	MON	0.2632750
9	LEU	0.3592121

Tak jak wspomniano wyżej, tworzone są wykresy z opisem danych (boxploty) oraz wykresy rozkładu danych (density plot), które są wyświetlane w zakładce Plots, ale także zapisywane do pliku. Przykładowe wykresy:



Boxplot oraz wykres gęstości

Białe kropki na boxplocie oznaczają wartości odstające. Im mniej ich jest, tym dane są bardziej jednolite i spójne. „Wąsy" pudełek oznaczają wartości minimalne i maksymalne, natomiast pogrubiona czarna linia wewnątrz pudełek oznacza średnią dla grupy.
Rozkład normalny na wykresie gęstości zbliżony jest do kształtu dzwonu (tzw. krzywa dzwonowa), dlatego im bardziej wykres przypomina ten kształt, tym bardziej jest zbliżony do rozkładu normalnego (Gaussa). Można powiedzieć, że na przykładowych wykresach, dane są średnio zbliżone do wykresu normalnego.

3) Analiza porównawcza pomiędzy grupami.

Parametry zgodnie z wynikami testów na rozkład i wariancję zostają podzielone na stosowne wobec nich testy porównujące grupy niezależne według poniższego schematu.

Schemat wyboru testu statystycznego

Ilość porównywanych grup	Zgodność z rozkładem normalnym	Jednorodność wariancji	Wybrany test
>2	TAK	TAK	Test ANOVA (*post hoc* Tukeya)
		NIE	Test Kruskala–Wallisa (*post hoc* Dunna)
	NIE	–	

Pierwsze zostają wyliczane testy ANOVA. Jeśli pvalue (Pr(>F)) tego testu jest wyższe, bądź równe 0.05, to nie ma istotnych różnic pomiędzy grupami. Jeśli jednak jest niższe niż 0.05, to należy wykonać test post hoc, w tym przypadku Tukeya. Wyświetlony wynik testu Tukeya ma następujące parametry:

- diff – różnica między średnimi wartościami dwóch porównywanych grup, im większa wartość tym większa różnica;
- lwr i upr – dolna i górna granica przedziału ufności dla różnicy między grupami (domyślnie 95%);
- p adj – skorygowana wartość pvalue gdzie wartość < 0.05 wskazuje na istotną różnicę między danymi grupami.

```
*********************************************************
Pr(>F) = 0.205627845266204
 Brak istotnych różnic w kolumnie wiek znalezione przy użyciu testu ANOVA.
*********************************************************
0.00185981027219031 < 0.05 - są różnice między grupami w kolumnie MCHC
                    diff         lwr      upr       p adj
CHOR1-KONTROLA 0.7261892 -0.02444453 1.476823 0.060043259
CHOR2-KONTROLA 1.1494120  0.39877827 1.900046 0.001352322
CHOR2-CHOR1    0.4232228 -0.32741093 1.173857 0.372940393
Wartosc p adj < 0.05 oznacza istotna roznice pomiedzy podanymi grupami
 Wykonano testem Tukeya.
*********************************************************
Pr(>F) = 0.596500941384286
 Brak istotnych różnic w kolumnie LEU znalezione przy użyciu testu ANOVA.
>
```

Drugie zostają wyliczane testy Kruskala–Wallisa. Jeśli p-value tego testu jest wyższe, bądź równe 0.05, to nie ma istotnych różnic pomiędzy grupami. Jeśli jednak jest niższe niż 0.05, to należy wykonać test post hoc – test Dunna. Wyświetlony wynik testu Dunna ma następujące parametry:

- Z – statystyka dla porównania par grup. Wartość ta pochodzi z rozkładu normalnego i wskazuje, jak daleko od średniej znajduje się obserwacja w jednostkach odchylenia standardowego, im większa wartość tym większa różnica;
- P.unadj – nieskorygowana wartość p-value dla porównania par grup;
- P.adj – skorygowana wartość p-value metodą Holma (domyślnie) dla porównania par grup, p adj < 0.05 oznacza istotną różnicę między grupami.

```
 Brak istotnych różnic w kolumnie PLT znalezione przy użyciu testu Kruskala-Wallisa.
*****************************************************
0.000767331503178641 < 0.05 - są różnice między grupami w kolumnie HGB
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Holm method.

        Comparison          Z      P.unadj       P.adj
1    CHOR1 - CHOR2 -0.1201601 0.9043562952 0.904356295
2 CHOR1 - KONTROLA  3.2183431 0.0012893348 0.002578670
3 CHOR2 - KONTROLA  3.3385033 0.0008423104 0.002526931
Wartosc P.adj < 0.05 oznacza istotna roznice pomiedzy podanymi grupami.
 Wykonano testem Dunna.
*****************************************************
0.0189609134312322 < 0.05 - są różnice między grupami w kolumnie HCT
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Holm method.

        Comparison         Z     P.unadj      P.adj
1    CHOR1 - CHOR2 0.7885475 0.430376497 0.43037650
2 CHOR1 - KONTROLA 2.7355785 0.006227073 0.01868122
3 CHOR2 - KONTROLA 1.9470310 0.051531020 0.10306204
Wartosc P.adj < 0.05 oznacza istotna roznice pomiedzy podanymi grupami.
 Wykonano testem Dunna.
*****************************************************
P-value = 0.25421349117813
 Brak istotnych różnic w kolumnie MON znalezione przy użyciu testu Kruskala-Wallisa.
```

4) Analiza korelacji pomiędzy parametrami w obrębie każdej grupy.

Parametryczny test korelacji Pearsona jest wykorzystywany do danych, które spełniają założenia normalności rozkładu i homogeniczności wariancji. W związku z tym test ten wykonywany jest na parametrach przechowywanych w liście z parametrami do testu ANOVA. Tym razem jednak testy korelacji są wykonywane między dwoma parametrami w obrębie tej samej grupy badawczej. Jeśli pvalue wyniku testu korelacji jest mniejsze od 0.05 to wskazuje to na korelację pomiędzy zmiennymi (parametrami, np. pomiędzy kolumnami danych *ERY* a *HGB*). Testy korelacji wykonywane są z każdym parametrem, który występuje w liście do testu ANOVA oraz dla każdej grupy.

Wyniki zestawiane i wyświetlane pod koniec segmentu są w tabeli *ramka_korelacji*, razem z wynikami testu korelacji Spearmana.

W przypadku parametrów nie spełniających założeń określonych dla testu Pearsona, wykonywany jest test korelacji Spearmana, który jest bardziej elastyczne pod względem założeń dotyczących danych, na parametrach przechowywanych w liście z parametrami do testu Kruskala–Wallisa. Schemat wyniku testu jest taki sam jak powyżej dla testu Pearsona.

We wspomnianej wyżej tabeli z zestawieniem wyników testów, znajdują się także określenia kierunku i siły korelacji na podstawie funkcji określonych dla współczynnika korelacji r.

Kierunek:

- $r > 0$ korelacja dodatnia – gdy zmienna X rośnie to Y także rośnie,
- $r = 0$ brak korelacji – gdy zmienna X rośnie to Y czasem rośnie a czasem maleje,
- $r < 0$ korelacja ujemna – gdy zmienna X rośnie to Y maleje.

Siła:

- $-1 < r \leq -0.7$ 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

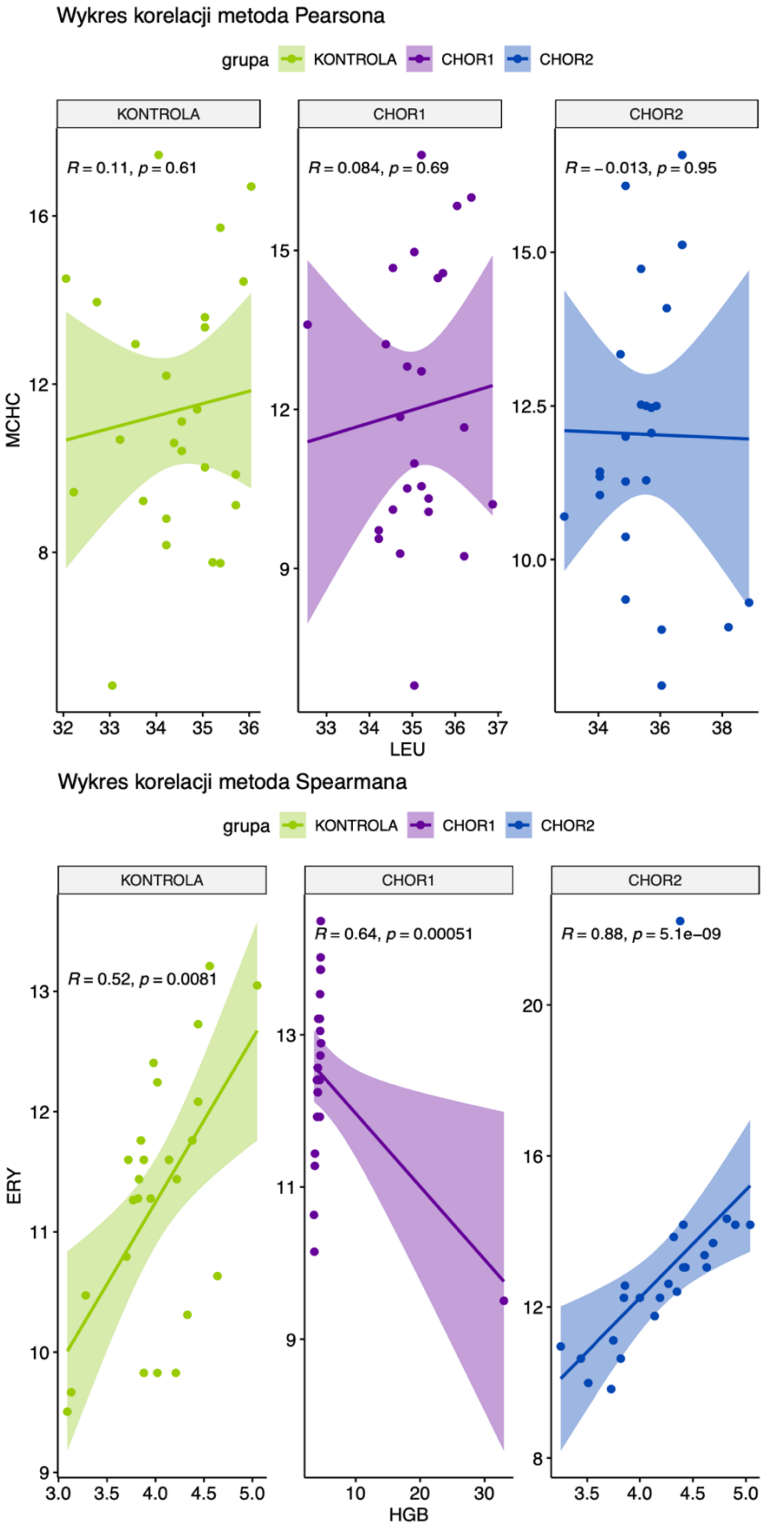
Schemat tabeli z wynikami testów korelacji:

	grupa	porownywana_para	p.value	r	korelacja	sila_korelacji	metoda
1	CHOR1	wiek + MCHC	7.922520e-02	0.0000000	brak	brak	Pearsona
2	CHOR1	wiek + LEU	8.309994e-01	0.0000000	brak	brak	Pearsona
3	CHOR1	MCHC + LEU	6.911065e-01	0.0000000	brak	brak	Pearsona
4	CHOR2	wiek + MCHC	8.495619e-01	0.0000000	brak	brak	Pearsona
5	CHOR2	wiek + LEU	1.670266e-01	0.0000000	brak	brak	Pearsona
6	CHOR2	MCHC + LEU	9.526777e-01	0.0000000	brak	brak	Pearsona
7	KONTROLA	wiek + MCHC	1.005930e-01	0.0000000	brak	brak	Pearsona
cor	KONTROLA	wiek + LEU	1.165459e-02	-0.4961399	ujemna	srednia ujemna	Pearsona
11	KONTROLA	MCHC + LEU	6.102397e-01	0.0000000	brak	brak	Pearsona
12	CHOR1	hsCRP + ERY	4.131638e-01	0.0000000	brak	brak	Spearmana
13	CHOR1	hsCRP + PLT	7.810855e-01	0.0000000	brak	brak	Spearmana
14	CHOR1	hsCRP + HGB	5.631822e-01	0.0000000	brak	brak	Spearmana

Tworzona jest także tabela przefiltrowana z rzeczywiście istotnymi korelacjami (*istotne_korelacje*).

	grupa	porownywana_para	p.value	r	korelacja	sila_korelacji	metoda
cor	KONTROLA	wiek + LEU	1.165459e-02	-0.4961399	ujemna	srednia ujemna	Pearsona
rho	CHOR1	ERY + HGB	5.088237e-04	0.6443040	dodatnia	silna dodatnia	Spearmana
rho1	CHOR1	ERY + HCT	1.068964e-03	0.6150144	dodatnia	silna dodatnia	Spearmana
rho2	CHOR1	HGB + HCT	6.306757e-12	0.9361754	dodatnia	bardzo silna dodatnia	Spearmana
rho3	CHOR2	ERY + HGB	5.085153e-09	0.8831042	dodatnia	bardzo silna dodatnia	Spearmana
rho4	CHOR2	ERY + HCT	1.493868e-05	0.7514456	dodatnia	bardzo silna dodatnia	Spearmana
rho5	CHOR2	HGB + HCT	2.648767e-05	0.7369646	dodatnia	bardzo silna dodatnia	Spearmana
rho6	KONTROLA	ERY + PLT	2.253247e-02	0.4542830	dodatnia	srednia dodatnia	Spearmana
rho7	KONTROLA	ERY + HGB	8.056956e-03	0.5175479	dodatnia	silna dodatnia	Spearmana
rho8	KONTROLA	ERY + HCT	1.916221e-04	0.6787296	dodatnia	silna dodatnia	Spearmana
rho9	KONTROLA	HGB + HCT	1.221057e-10	0.9167162	dodatnia	bardzo silna dodatnia	Spearmana

Podczas testowania korelacji, tworzone są także wykresy korelacji dla każdej pary parametrów i grupy wraz z dodanym trendem regresji liniowej. Można je wyświetlić po wejściu w plik *Wykresy-korelacja.pdf* zapisany w katalogu roboczym.



Dane na wykresach korelacji nie zbierają się wokół linii, więc można stwierdzić, że w tych przykładowych danych jest dosyć niska korelacja. Jedynym wyjątkiem jest grupa CHOR2 na drugim zestawie wykresów metodą Spearmana, gdzie korelacja jest b. silna, r = 0.88, co i widać po wykresie, gdzie punkty zbierają się wokół linii.

Podsumowanie

Powyższe narzędzie służy do podstawowej analizy danych medycznych, które może pomóc w dalszej, głębszej analizie poszczególnych parametrów badanych na grupach pacjentów. Użyty język R jest jednym z najczęściej używanych narzędzi w analityce danych biologicznych, chemicznych oraz medycznych, ponieważ można sprawnie wykonywać działania na danych w sposób kontrolowany i pożądany.