Statystyczna analiza danych – Projekt Podstawowa analiza statystyczna dla danych medycznych

Joanna Wełnic, indeks: 155657, gr. 1, semestr 4, Bioinformatyka

Narzędzie w formie skryptu w języku R zostało stworzone w celu przeprowadzenia podstawowej analizy statystycznej dla danych medycznych. Narzędzie to zostało podzielone na cztery etapy, które umożliwiają kompleksową analizę danych i uzyskanie istotnych informacji z zestawów danych medycznych.

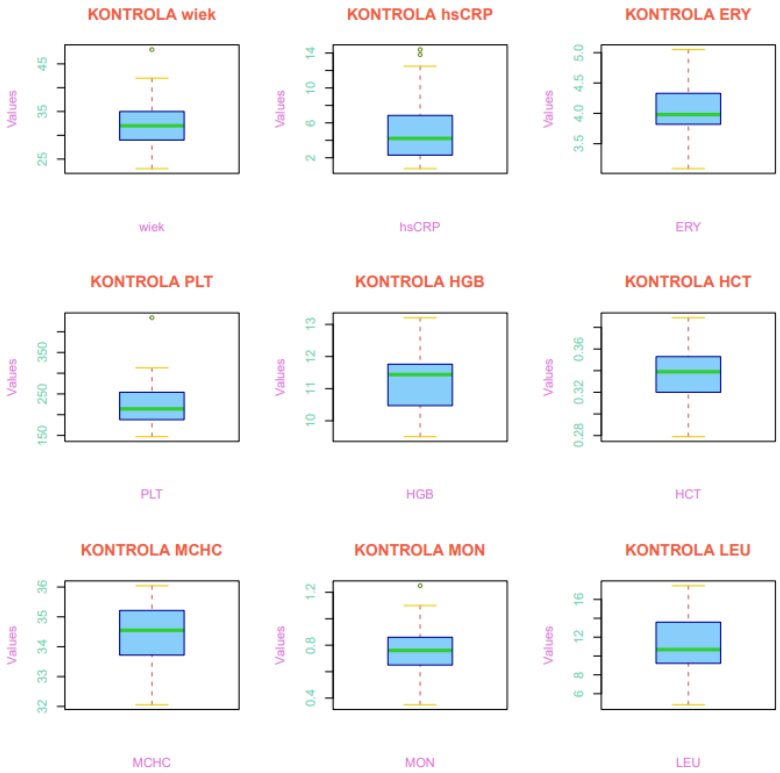
Pierwszy etap:

Funkcja `unique()` została użyta do uzyskania nazw poszczególnych grup znajdujących się w bazie danych, a następnie te informacje zapisano do zmiennej. Wyodrębnione zostały również nazwy kolumn zawierających wartości numeryczne, które mogły wymagać uzupełnienia brakujących danych średnią z danej grupy. Utworzono funkcję pobierającą dane dla danej grupy i kolumny wymagające uzupełnienia. Dla każdej znalezionej grupy, wyodrębnione dane zostały przekazane do funkcji, a następnie zaktualizowano oryginalną ramkę danych. Zaraportowanie wszystkich wprowadzonych zmian możemy znaleźć w konsoli:

```
Wprowadzono średnią wartość 12.4114125 dla kolumny HGB w grupie CHOR1 w wierszach: 13
Wprowadzono średnią wartość 0.857916666666667 dla kolumny MON w grupie CHOR1 w wierszach: 5
Wprowadzono średnią wartość 11.263575 dla kolumny HGB w grupie KONTROLA w wierszach: 68
```

Następną rzeczą jest stworzenie funkcji, która raportuje wartości odstające. W niej tworzą się boxploty (zawierające m.in. informacje na temat mediany, kwartyli, „wąsów” i wartości odstających; każda część boxplota jest też pokolorowana na indywidualny kolor w celu wyróżnienia). Następnie dla każdej grupy zapisywane są one do osobnego pliku .pdf, w którym znajdują się utworzone wykresy. Układ wykresów jest dynamiczny, przez co nie wystąpią problemy przy różnej ilości kolumn. Dodatkowo utworzony raport o wartościach odstających możemy znaleźć w plikach *_outliers.txt dla każdej grupy, jeśli będziemy chcieli odczytać wartości o większej dokładności niż na samym wykresie.

Przykładowe utworzone boxploty dla grupy KONTROLA:



Kropki na boxplocie oznaczają wartości odstające, im mniej ich jest tym dane są bardziej jednolite. W centrum każdego boxplotu znajduje się linia oznaczająca medianę, czyli wartość środkową zbioru danych. Długość pudełka wskazuje na zakres wartości, w którym znajduje się większość danych, z dolnym i górnym końcem reprezentującymi pierwszy i trzeci kwartyl. Wąsy na końcach boxplotu przedłużają się w stronę skrajnych wartości, ale niezbyt daleko od obszaru, który zawiera większość danych. Są one używane do wychwycenia potencjalnych wartości odstających, czyli danych znacznie różniących się od reszty zbioru.

Przykładowa zawartość pliku zawierającego raport z wartościami odstającymi o nazwie KONTROLA_outliers.txt:

```
Grupa: KONTROLA Kolumna: wiek Wartości odstające: 48
Grupa: KONTROLA Kolumna: hsCRP Wartości odstające: 14.3951, 13.8106
Grupa: KONTROLA Kolumna: PLT Wartości odstające: 434
Grupa: KONTROLA Kolumna: MON Wartości odstające: 1.25
```

Dzięki tej informacji wiemy, które dane wpływają na niejednolitość danych. Możemy np. wziąć je pod uwagę w dalszej analizie lub wyłączyć z analizy.

Drugi etap:

Wykonane zostały charakterystyki dla badanych grup w formie podsumowania za pomocą m.in. funkcji group_by_at() oraz summarise(). Dzięki temu zostały pozyskane dane tj. minimalna wartość, mediana, średnia, maksymalna wartość, odchylenie standardowe, rozstęp międzykwarylowy i wariancja w próbie dla każdej kolumny z wartościami numerycznymi. Te informacje przedstawione są w osobnych wierszach dla każdej grupy.

Wynik jest zapisywany do pliku podsumowanie_dane.csv. Przykładowy wycinek tabeli:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	grupa	wiek_count	wiek_min	wiek_media	wiek_mean	wiek_max	wiek_sd	wiek_IQR	wiek_var	hsCRP_cour	hsCRP_min	hsCRP_med	hsCRP_mea	hsCRP_max
2	CHOR1	25	17	29	29,56	43	5,88	6	34,59	25	0,49	3,97	6,1	42,65
3	CHOR2	25	21	30	30,04	42	5,9	8	34,79	25	0,34	3,45	5,54	19,21
4	KONTROLA	25	23	32	32,32	48	5,61	6	31,48	25	0,76	4,22	5,3	14,4

W pliku "podsumowanie_wykresy.pdf" znajduje się graficzne zestawienie kluczowych charakterystyk danych w sposób zwięzły i czytelny.

Wykresy słupkowe odchylenia standardowego dla każdej grupy - pozwalają one na szybką wizualizację rozproszenia danych wokół średniej w każdej z grup. Im wyższy słup, tym większe odchylenie standardowe i większa zmienność danych w danej grupie.

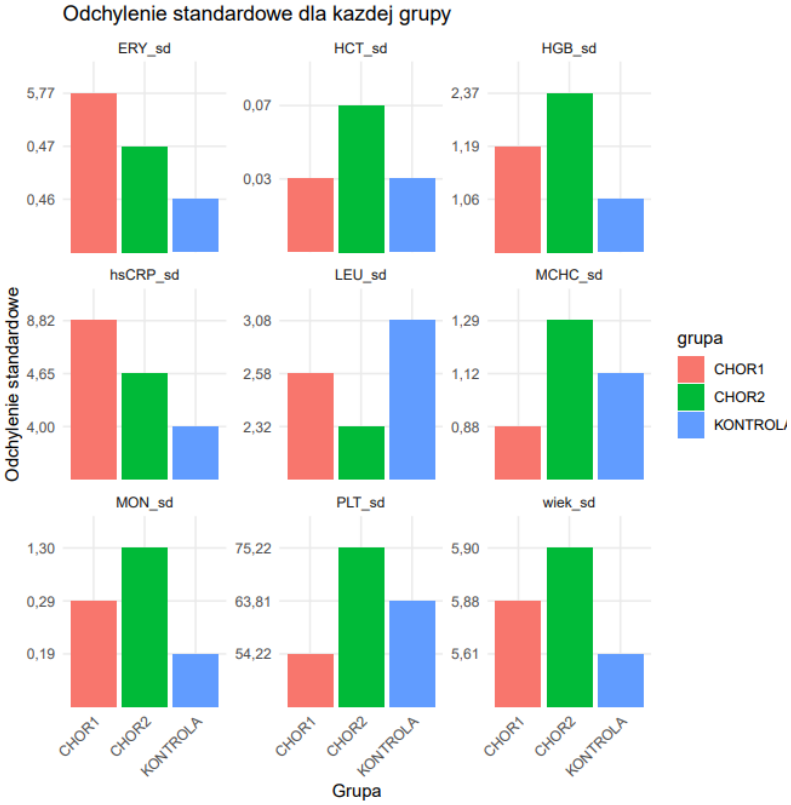
Rozstęp międzykwartylowy dla każdej grupy - przedstawiają różnicę między pierwszym a trzecim kwartylem w każdej grupie. Jest to miara zmienności danych, która uwzględnia tylko wartości wewnętrzne, eliminując wpływ wartości odstających na wynik.

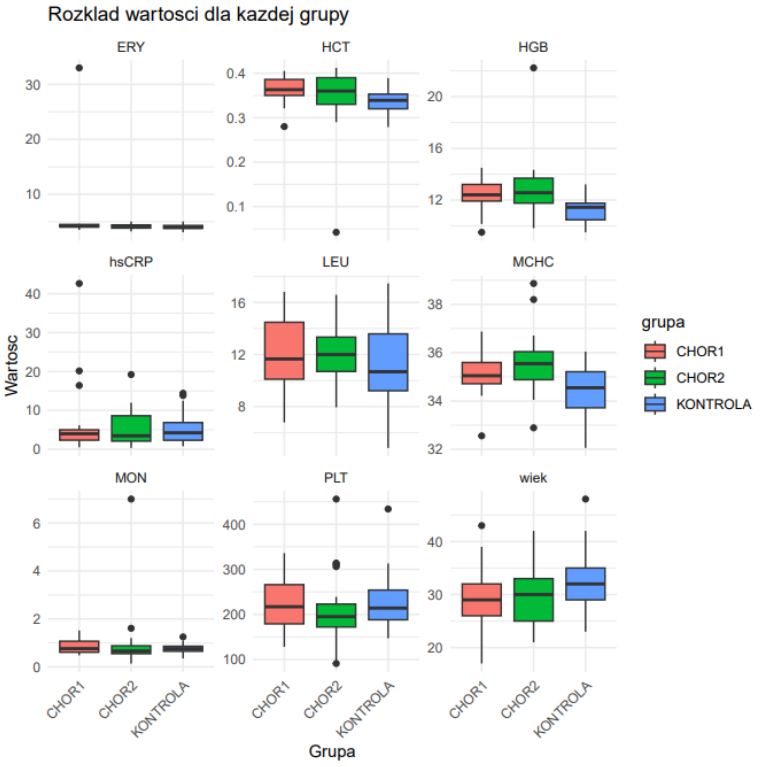
Wariancja dla każdej grupy - pokazuje, jak bardzo wartości w danej grupie różnią się od średniej. Im większa wariancja, tym większa zmienność danych w danej grupie.

Rozkład wartości dla każdej grupy za pomocą boxplotów - umożliwiają one porównanie rozkładów danych między grupami. Linia wewnątrz pudełka to mediana, a długość pudełka pokazuje zakres, w którym znajduje się większość danych. Wąsy wskazują na obszar, w którym znajdują się dane nieodstające.

Te wykresy są przydatne do analizy danych, pomagając zidentyfikować istotne różnice między grupami i próbami oraz zrozumieć zmienność danych w badanych grupach.

Przykładowe wykresy znajdujące się w pliku podsumowanie_wykresy.pdf:





Trzeci etap:

Analizy porównawcze zostały wykonane, a istotne statystycznie różnice pomiędzy grupami zostały zidentyfikowane i zawarte w raportach. Zaraportowanie o istotnych statystycznie różnicach, pomiędzy którymi grupami występują, wartość p-value, nazwa kolumny (z zawartością numeryczną) oraz istotność różnicy (przedstawione za pomocą różnicy 0.05 i p-value) znajduje się w dwóch plikach:

- Raport z testu ANOVA został zapisany w pliku "tukey_report.txt". W raporcie tym znajdziesz szczegółowe informacje na temat istotnych różnic pomiędzy grupami, wraz z nazwami kolumn, wartościami p-value oraz istotnością różnicy.
- Raport z testu Kruskala-Wallisa został zapisany w pliku "dunn_raport.txt". W tym raporcie również zawarte są istotne statystycznie różnice pomiędzy grupami, wraz z nazwami kolumn, wartościami p-value oraz istotnością różnicy.

Wszystkie testy zostały przeprowadzone zgodnie z tabelą:

Tablica 1: Wyboru testu statystycznego dla 2 i > 2 grup niezależnych.

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	-	

Wykonanywane testy dotyczą grup niezależnych. Wybór testów parametrycznych i nieparametrycznych dla danych zależy od powyższej tabeli.

Wykonywane są testy zgodności z rozkładem normalnym - Shapiro oraz na jednorodność wariancji - Levene w celu wybrania odpowiednich następnych testów – Anova bądź Kruskala-Wallisa.

Przykładowa zawartość pliku tukey_report.txt:

porownanie_grup	nazwa_kolumny	diff	lwr	upr	p adj	istotnosc_roznicy	Info
KONTROLA-CHOR2	MCHC	-1.149412	-1.90004573279819	-0.39877826720181	0.00135232152206555	0.0486476784779345	Istotne statystycznie wartości p-value

Przykładowa zawartość pliku dunn_raport.txt:

Comparison	Z	P.unadj	P.adj	istotnosc_roznicy	Info	nazwa_kolumny	
CHOR1 - KONTROLA	3.21834312269376	0.00128933478647409	0.00193400217971114	0.0480659978202889	Istotne statystycznie wartości p-value	HGB	
CHOR2 - KONTROLA	3.33850325946437	0.00084231042109912 4	0.00252693126329737	0.0474730687367026	Istotne statystycznie wartości p-value	HGB	
CHOR1 - KONTROLA	2.73557853518423	0.00622707251359907	0.0186812175407972	0.0313187824592028	Istotne statystycznie wartości p-value	HCT	

W trakcie wykonywania odpowiednich testów w konsoli wyświetla się raport z ich przeprowadzania, przykładowo:

```
Test Shapiro-Wilka dla kolumny: MON
# A tibble: 3 x 3
  grupa    statistic       p.value
  <fct>        <dbl>         <dbl>
1 CHOR1        0.910 0.0311
2 CHOR2        0.410 0.000000000539
3 KONTROLA     0.959 0.396
Test Kruskala-Wallisa dla kolumny: MON
0.2542135 > 0.05 - brak różnic pomiędzy grupami
Test Shapiro-Wilka dla kolumny: LEU
# A tibble: 3 x 3
  grupa    statistic p.value
  <fct>        <dbl>   <dbl>
1 CHOR1        0.955   0.331
2 CHOR2        0.969   0.615
3 KONTROLA     0.980   0.877
Test Levene'a dla kolumny: LEU
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  1.0385 0.3592
      72
Test ANOVA dla kolumny: LEU
0.5965009 > 0.05 - brak różnic pomiędzy grupami
```

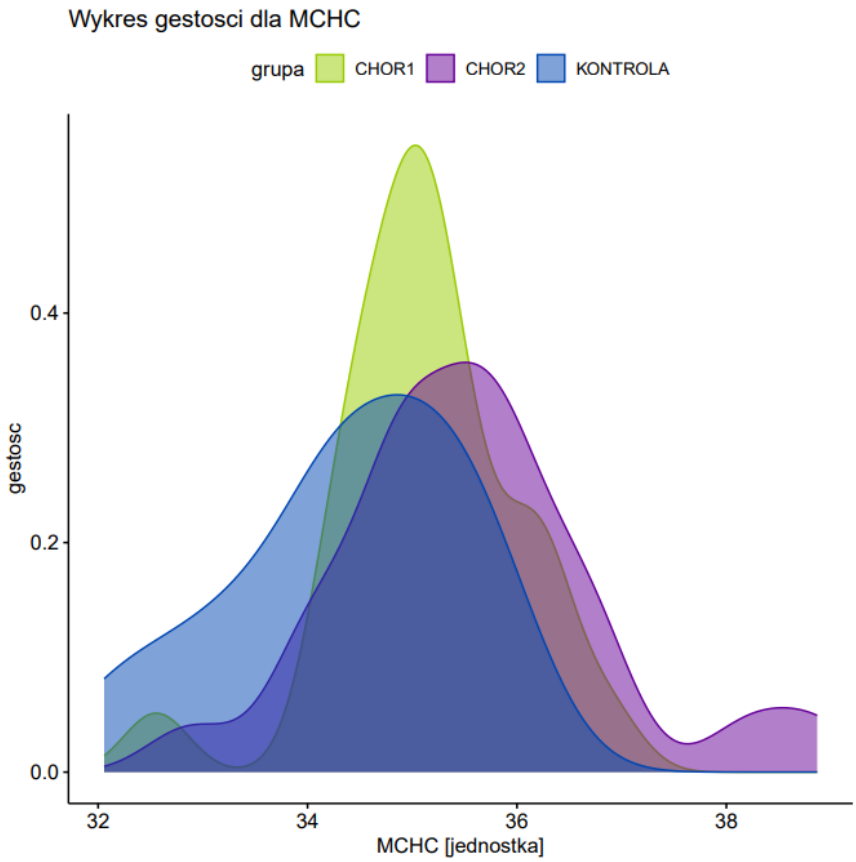
Wykresy gęstości graficznie ilustrujące zgodność z rozkładem normalnym mogą być interpretowane jako narzędzie do wizualnej oceny rozkładu danych w poszczególnych grupach badanej próby. Poprzez nakładanie się gęstości rozkładów dla różnych grup na jednym wykresie, można łatwo porównać kształt i rozkład wartości w poszczególnych grupach.

Interpretacja takich wykresów może obejmować:

1. **Porównanie kształtu rozkładów**: Nakładanie się krzywych gęstości dla różnych grup pozwala ocenić, czy rozkłady danych w poszczególnych grupach są podobne czy różnią się między sobą. Jeśli krzywe gęstości są zbliżone do siebie, sugeruje to podobny kształt rozkładu w badanych grupach.
2. **Identyfikacja odstępstw od rozkładu normalnego**: Jeśli krzywe gęstości dla różnych grup różnią się znacząco od rozkładu normalnego (np. są asymetryczne lub mają dodatkowe szczyty), może to wskazywać na odstępstwa od założenia normalności w danych.
3. **Detekcja nietypowych zachowań**: Gęstość rozkładu może ujawnić nietypowe zachowania w danych, takie jak skupienie się wartości w określonych przedziałach lub występowanie długich ogonów.
4. **Podkreślenie istotnych różnic**: Jeśli krzywe gęstości dla różnych grup znacząco się różnią, może to wskazywać na istotne różnice między grupami w rozkładzie badanej cechy.

Wykresy gęstości graficznie ukazujące zgodność z rozkładem normalnym znajdują się w pliku gestosc_wykresy.pdf.

Przykładowy wykres z pliku gestosc_wykresy.pdf:



Powyższy przykładowy wykres dla MCHC dla grup CHOR1, CHOR2 i KONTROLA przypomina krzywą Gaussa, która ma charakterystyczny kształt dzwonu, gdzie większość obserwacji koncentruje się wokół środka (wartości średniej), a odchylenia standardowe określają szerokość i symetrię dzwonu. Dlatego ten wykres gęstości danych, sugeruje, że badane dane mogą być zbliżone do rozkładu normalnego. Szczególne nakładanie się wykresów możemy zaobserwować dla grup CHOR2 i KONTROLA. CHOR1 zauważalnie odstaje w górę, jednak wciąż możemy zaobserwować spore nałożenie się z resztą.

Czwarty etap:

Przeprowadzono analizę korelacji i zaraportowano istotne statystycznie korelacje pomiędzy parametrami oraz w obrębie określonych grup. Określono siłę i kierunek korelacji. Dla danych parametrycznych zastosowano test Pearsona, a dla danych nieparametrycznych – test Spearmana. Podsumowane dane, obejmujące jedynie istotne statystycznie korelacje (p-value < 0.05), zostały zapisane w formie tabeli. Tabela zawiera informacje na temat dodatnich, ujemnych lub braku korelacji. W R Studio dane te są dostępne w tabeli "results":

grupa	porownywana_para	p_value	r	korelacja	sila_korelacji	metoda
CHOR1	ERY-HGB	5.088237e-04	0.6443040	korelacja dodatnia	silna korelacja dodatnia	spearman
CHOR1	ERY-HCT	1.068964e-03	0.6150144	korelacja dodatnia	silna korelacja dodatnia	spearman
CHOR1	HGB-HCT	1.870428e-13	0.9533424	korelacja dodatnia	bardzo silna korelacja dodatnia	pearson
CHOR1	HGB-MCHC	6.117491e-04	0.6373059	korelacja dodatnia	silna korelacja dodatnia	pearson
CHOR1	HCT-MCHC	3.335169e-02	0.4268167	korelacja dodatnia	korelacja dodatnia o średnim natężeniu	pearson
CHOR2	ERY-HGB	5.085153e-09	0.8831042	korelacja dodatnia	bardzo silna korelacja dodatnia	spearman
CHOR2	ERY-HCT	1.493868e-05	0.7514456	korelacja dodatnia	bardzo silna korelacja dodatnia	spearman
CHOR2	PLT-MCHC	7.798945e-03	-0.5193725	korelacja ujemna	silna korelacja ujemna	spearman
CHOR2	HGB-HCT	2.648767e-05	0.7369646	korelacja dodatnia	bardzo silna korelacja dodatnia	spearman
KONTROLA	wiek-ERY	2.037864e-02	0.4609923	korelacja dodatnia	korelacja dodatnia o średnim natężeniu	pearson
KONTROLA	wiek-LEU	1.165459e-02	-0.4961399	korelacja ujemna	korelacja ujemna o średnim natężeniu	pearson
KONTROLA	hsCRP-LEU	3.509101e-02	0.4253846	korelacja dodatnia	korelacja dodatnia o średnim natężeniu	spearman
KONTROLA	ERY-PLT	2.253247e-02	0.4542830	korelacja dodatnia	korelacja dodatnia o średnim natężeniu	spearman
KONTROLA	ERY-HGB	2.032562e-03	0.5871174	korelacja dodatnia	silna korelacja dodatnia	pearson
KONTROLA	ERY-HCT	5.223601e-06	0.7757054	korelacja dodatnia	bardzo silna korelacja dodatnia	pearson
KONTROLA	ERY-MCHC	1.206322e-02	-0.4940718	korelacja ujemna	korelacja ujemna o średnim natężeniu	pearson
KONTROLA	HGB-HCT	2.244245e-12	0.9418013	korelacja dodatnia	bardzo silna korelacja dodatnia	pearson
KONTROLA	MON-LEU	7.736271e-03	0.5198233	korelacja dodatnia	silna korelacja dodatnia	pearson

Tabela zapisuje się również automatycznie do pliku korelacja_report.txt.

Ze współczynnika korelacji r odczytano zależności:

• 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.

Natomiast siła korelacji jest opisana zgodnie z poniższymi przedziałami:

• $-1 < r \leq -0.7$ 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

Wykresy korelacji zostały podzielone na dwie grupy:

Wykresy dla korelacji Pearsona z regresją liniową (przykład po lewej stronie):

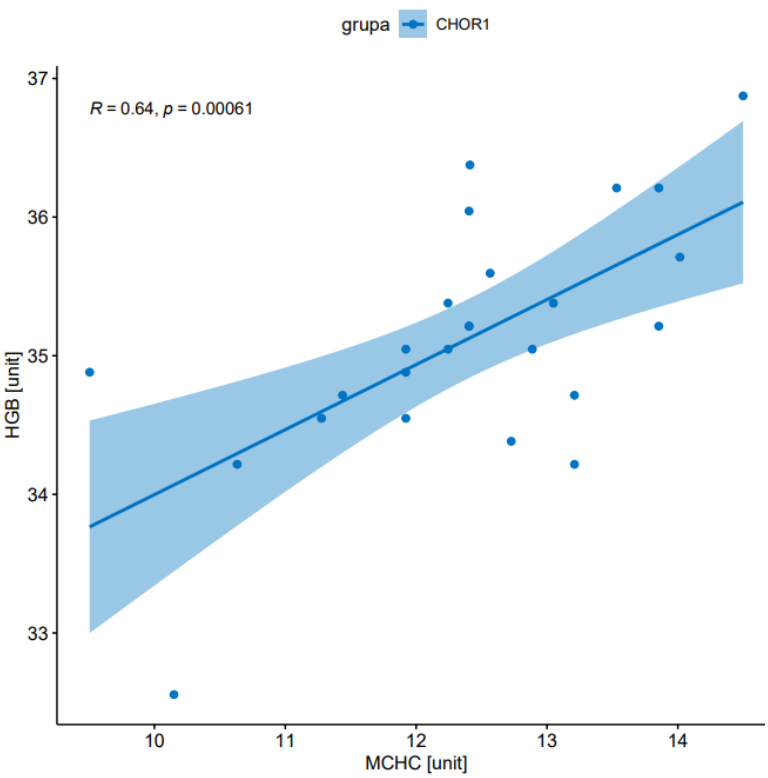
- **Regresja liniowa**: Linia regresji wskazuje na trend lub zależność liniową między zmiennymi. Kierunek i nachylenie linii wskazują na siłę i kierunek tej zależności. Im bardziej nachylona linia, tym silniejsza korelacja między zmiennymi. Jeśli linia jest dodatnia, oznacza to, że zmienne wzrastają razem. Jeśli jest ujemna, zmienne maleją razem.
- **Rozrzut punktów wokół linii regresji**: Rozrzut punktów wokół linii regresji wskazuje na stopień dopasowania modelu do danych. Im mniejszy rozrzut, tym lepiej model liniowy pasuje do danych. Duży rozrzut może wskazywać na obecność nieliniowych związków lub obserwacje odstające.

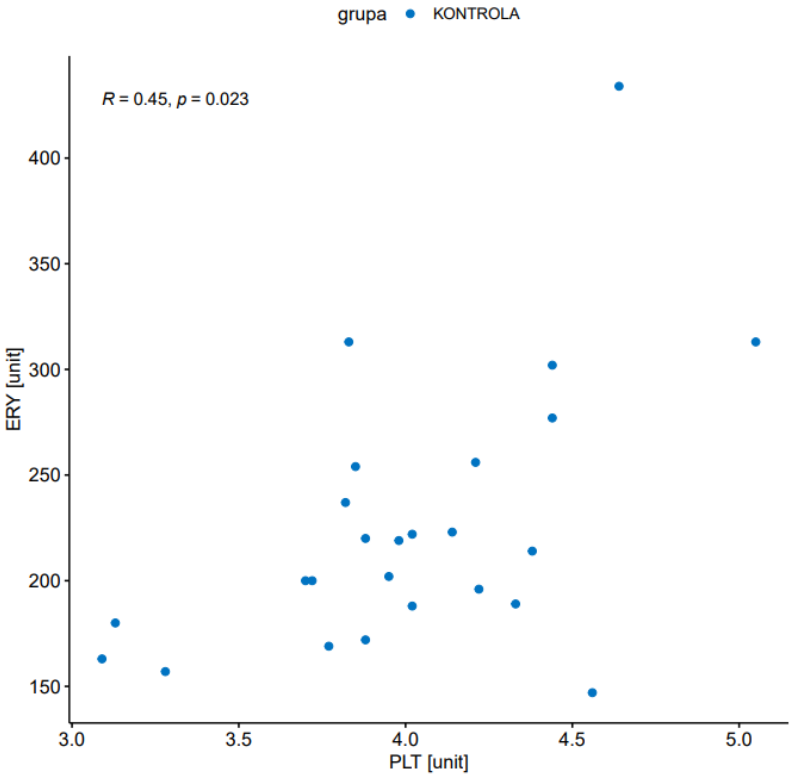
Wykresy dla korelacji Spearmana bez regresji liniowej (przykład po prawej stronie):

- **Zależność monotoniczna**: Brak linii regresji oznacza, że korelacja Spearmana opiera się na porządku rang, a nie na liniowej zależności. Wykres pokazuje kierunek i siłę monotonicznej zależności między zmiennymi. Jeśli punkty na wykresie układają się wzdłuż linii prostej, to sugeruje silną zależność monotoniczną między zmiennymi.
- **Sensory**: Wzrost wartości jednej zmiennej wraz ze wzrostem rangi drugiej zmiennej lub malejąca tendencja wskazuje na zależność negatywną. Brak wyraźnego wzrostu lub spadku sugeruje brak związku lub niemonotoniczną zależność.

Graficznie przedstawiona korelacja pomiędzy parametrami została zapisana do pliku korelacja_wykresy.pdf.

Przykładowa zawartość pliku korelacja_wykresy.pdf:





Na powyższym przykładowym wykresie po lewej możemy zaobserwować dodatnią korelację pomiędzy HGB i MCHC dla grupy CHOR1. Wartość współczynnika korelacji R=0.64 oznacza silną dodatnią korelację, co sugeruje, że gdy wartość jednej zmiennej rośnie, wartość drugiej zmiennej również ma tendencję do wzrostu. Wartość p=0.00061 jest bardzo niska (znacznie poniżej poziomu istotności 0.05), co wskazuje na to, że wynik korelacji jest statystycznie istotny. Oznacza to, że istnieje bardzo małe prawdopodobieństwo, że obserwowany związek jest wynikiem przypadku.

Na wykresie po prawej natomiast dotyczącym grupy KONTROLA widzimy umiarkowaną dodatnią korelację. To oznacza, że istnieje zauważalny związek między zmiennymi, ale nie jest on bardzo silny. Wartość p oznacza, że wynik jest statystycznie istotny w tym przypadku.