Sprawozdanie

Anna Szymczak 158070

1.Wstęp

Skrypt R ma posłużyć do wstępnej do wstępnej analizy statystycznej danych medycznych, które są zamieszczone w pliku.csv

Analiza danych:

1. Przygotowanie danych wejściowych
2. charakterystyki dla badanych grup
3. Analiza porównawcza pomiędzy grupami
4. Analiza korelacji

Przeprowadzone testy

1.Sapiro -Wilk
2. ANOVA
3. Turkey
4.Kruskala-Wallisa
4. DunnTest
5.Leneve
6.Spearman

2.Opis działania

Przygotowanie danych wejściowych

Narzędzie w pierwszej kolejności zaczyna od wczytania danych medycznych z pliku csv. Plik jest wczytany za pomocą funkcji read.csv2, która umożliwia odczytanie pliku csv.

Plik został zapisany do zmiennej data.

```
data<-read.csv2("C:/Users/pc/Desktop/ProjektR/przykladoweDane-Projekt (2).csv")
```

Następnie skrypt sprawdza, gdzie znajdują się braki danych (NA) za pomocą funkcji which() aby znaleźć wszystkie komórki z wartościami NA. Zapisuje do zmnnej braki.

```
braki<- which(is.na(data),arr.ind = TRUE)
```

Następnie zostają utworzone grupy (CHOR1,CHOR2,KONTROLA) tworząc podgrupy danych o nazwie zmiennej data_group.

Indeksy komórek pod którymi znajdują się brakujące dane:

```
     row col
[1,]  13   7
[2,]  68   7
[3,]   5  10
```

Za opmocą pętli przechodzi po każdej grupie i sprawdza wartości NA w kolumnie jeżeli znajduja się to wprowadza średnią w miejsce pustej komórki. Wprowadzona średnia jest średnią wartości w kolumnie danej grupy. Uzupełnienie braków za pomocą funkcji impute()

```
for(i in 1:nrow(braki)){
  if(i==1){
    print("1 wartość to wiersz 2 wartość to kolumna")
  }
  cat("brakuje danych w tej komórce",braki[i,],"\n")
}

# utworzenie grup badanych (chor1 chor2 kontrolne)
data_group <- list()
for(group in unique(data[,1])){ # unikalne wartosci w kolumnach każdej grupy
  data_group[[length(data_group)+1]] <- data[data[,1] == group,]
}
for(i in 1:length(data_group)){ # przechodzi po grupach
  for(j in colnames(data)){
    if(any(is.na(data_group[[i]][[j]]))){ # znajduje wartości puste
      data_group[[i]][j] <- impute(data_group[[i]][j], mean) # przypisuje średnią z grupy w miejsca puste
    }
  }
  print(data_group[[i]])
}
```

charakterystyki dla badanych grup

Skrypt szuka wartości odstających przechodząc po kazdej grupie. Za pomocą funkcji boxplot.stats() oblicza statystyki dla danej grupy i kolumniy.

Sprawdza:

Min – minimalną wartość

Q1 – pierwszy kwartyl

Median – mediane

Q3 – trzeci kwartyl

Max – maksymalną wartość

Outliers – wartości odstające

Statystyki te zostają zapisane do pliku csv

```
# sprawdza wartości odstające
all_stats <- data.frame()
for(j in 1:length(data_group)){
  group_stats <- data.frame()
  numeric_columns <- list();
  for(i in 3:11){
    stats <- boxplot.stats(data_group[[j]][, i])
    # podsumowanie
    col_stats <- data.frame(
      Group = data_group[[j]][1,1],
      Variable = colnames(data)[i],
      Min = round(min(data_group[[j]][, i], na.rm = TRUE),digits = 2),
      Q1 = round(quantile(data_group[[j]][, i], 0.25, na.rm = TRUE),digits = 2),
      Median = round(median(data_group[[j]][, i], na.rm = TRUE),digits = 2),
      Q3 = round(quantile(data_group[[j]][, i], 0.75, na.rm = TRUE),digits = 2),
      Max = round(max(data_group[[j]][, i], na.rm = TRUE),digits = 2),
      Mean = round(mean(data_group[[j]][, i], na.rm = TRUE),digits = 2),
      Outliers = paste(round(stats$out, 2), collapse = ", ")
    )
    group_stats <- bind_rows(group_stats, col_stats)
  }
  all_stats <- bind_rows(all_stats, group_stats)
}

write.csv2(all_stats, file = "statystyki.csv", row.names = FALSE)
```

Łączenie statystyk dla wszystkich grup:

col_stats – zminenna przechowuje statystyki danej grupy kolumny

group_stats – zminna przechowuje statystyki wszystkie kolumny danej grupy

all_stats – zmienna przechowuje statystyki wszystkie grupy

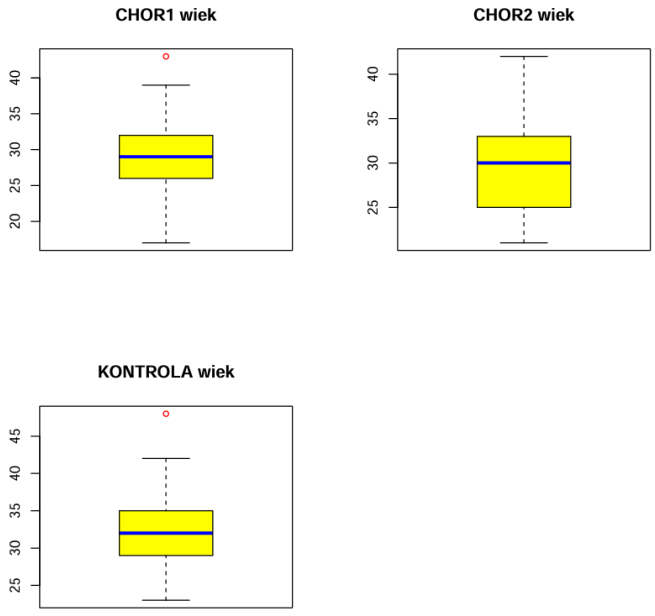
Dodanie col_stats do group_stats przy użyciu funkcji bind_rows(). Po wykonaniu dodaje wszystkie grupy group_stats do all_stats

Przykład zapisanych wyników:

Group	Variable	Min	Q1	Median	Q3	Max	Mean	Outliers
CHOR1	wiek	17	26	29	32	43	29,56	43
CHOR2	wiek	21	25	30	33	42	30,04	
KONTROLA	wiek	23	29	32	35	48	32,32	48
CHOR1	hsCRP	0,49	2,32	3,97	4,99	42,65	6,1	20.15, 16.41, 42.65
CHOR2	hsCRP	0,34	2,08	3,45	8,61	19,21	5,54	19.21
KONTROLA	hsCRP	0,76	2,3	4,22	6,85	14,4	5,3	14.4, 13.81
CHOR1	ERY	3,53	4,07	4,2	4,51	33	5,36	33
CHOR2	ERY	3,25	3,85	4,27	4,43	5,04	4,2	
KONTROLA	ERY	3,09	3,82	3,98	4,33	5,05	4,01	
CHOR1	PLT	128	179	217	266	336	225,28	
CHOR2	PLT	91	172	195	223	456	209,12	456, 314, 91, 311, 306
KONTROLA	PLT	147	188	214	254	434	225,88	434
CHOR1	HGB	9,5	11,92	12,4	13,21	14,5	12,41	0.9
CHOR2	HGB	9,83	11,76	12,57	13,69	22,23	12,81	22.23
KONTROLA	HGB	9,5	10,47	11,44	11,76	13,21	11,26	
CHOR1	HCT	0,28	0,35	0,36	0,39	0,41	0,36	0.28
CHOR2	HCT	0,04	0,33	0,36	0,39	0,41	0,35	0.04
KONTROLA	HCT	0,28	0,32	0,34	0,35	0,39	0,34	

Skrypt wykonane statystyki zapisuje do pliku pdf

Przykład



Analiza porównawcza pomiędzy grupami

Testy statystyczne zostały wykonane za pomocą ANOVA lub testu Kruskala-Wallisa.

Dane medyczne posiadają więcej niż dwie grupy badane.

Skrypt sprawdza normalość rozkładu za pomocą testu Shapiro-Wilka dla kaśdej grupy, gdy p-value jest mniejsze niż 0.05 zmienna normal_distribution jest ustawiona na 0, brak normalności.

Test ANOVA:

Jak zminenna ma normalny rozkład przeprowadzany jest test Levene'a na jednorodność wariancji. Jeśli test Levene'a potwierdzi jednorodność wariancji (p-value > 0.05), przeprowadzany jest test ANOVA. Gdy test ANOVA wykaże istotne różnice pomiędzy grupami zmienna jest dodawana do listy homogenious_relevant wraz z wynikami testu Tukeya.

Test Kruskala-Wallisa:

Jak zmienna nie spełnia warunku normalości , jednorodności wariancji wtedy jest przeprowadzany test Kruskala-Wallisa. Gdy test Kruskala-Walisa wykazuje istotne różnice między grupami zmienna jest dodawana do listy irregular_relevant wraz z wynikami testu Dunna.

Wyniki są zapisane do plików csv

Analiza korelacji

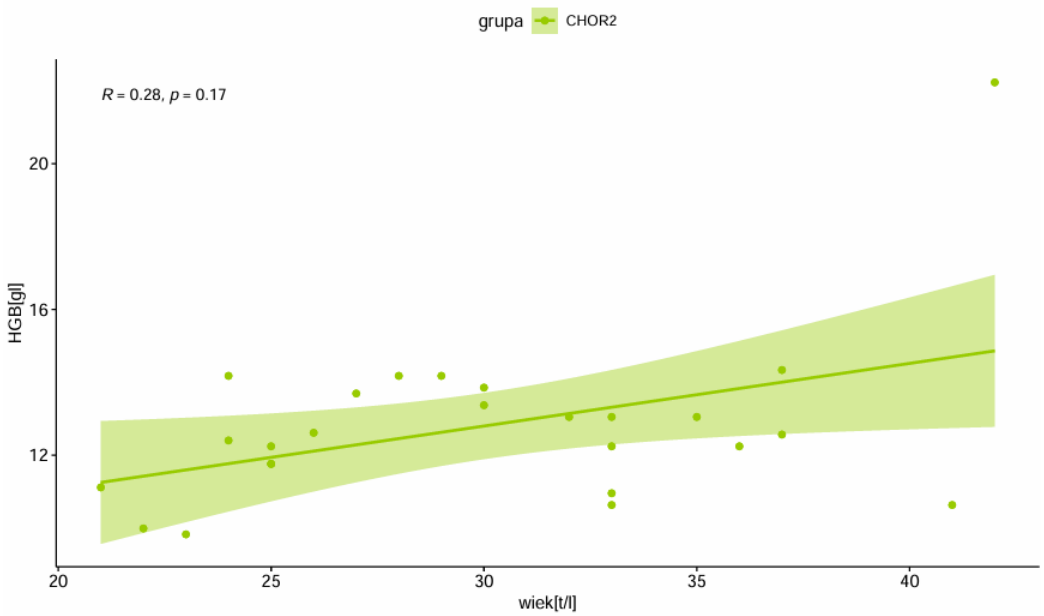
Skrypt analizuje korelację pomiedzy zmiennymi każdej grupy poprzez przypisywanie siły korelacji w zależności od wartości wsółczynnika korelacji.

test– współczynnik korelacji

test > 0.7 Bardzo mocna korelacja
test > 0.5 mocna dodatnia korelacja
test > 0.3 umiarkowana dodatnia korelacja
test > 0.2 słaba dodatnia korelacja
test > -0.2 brak korelacji
test > -0.3 słaba negatywna korelacja
test > -0.5 umiarkowana dodatnia korelacja
test > -0.7 silna negatywna korelacja

Zapisuje do pliku pdf w formie graficznej przedstwienie korelacji. Utworzenie wykresów wykonał za pomocą funkcji ggcatter().

Przykład wykresu:



3.Analiza wyników

Zapisane ogólne statystyki do tabeli dla każdej grup. W kolumnach znajdują się nazwy Variable, Min, Q1, Median, Q3, Max, Mean , Outliers sa to parametry statystyczne które opisują daną komórkę. W wierszach znajdują się parametry dla każdej grupy.

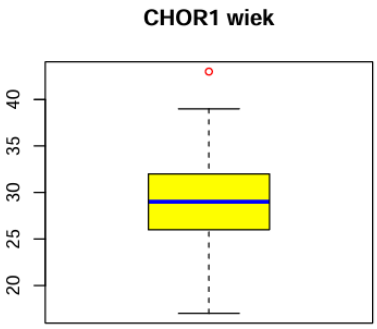
Group	Variable	Min	Q1	Median	Q3	Max	Mean	Outliers
CHOR1	wiek	17	26	29	32	43	29,56	43
CHOR2	wiek	21	25	30	33	42	30,04	
KONTROLA	wiek	23	29	32	35	48	32,32	48
CHOR1	hsCRP	0,49	2,32	3,97	4,99	42,65	6,1	20.15, 16.41, 42.65
CHOR2	hsCRP	0,34	2,08	3,45	8,61	19,21	5,54	19.21
KONTROLA	hsCRP	0,76	2,3	4,22	6,85	14,4	5,3	14.4, 13.81
CHOR1	ERY	3,53	4,07	4,2	4,51	33	5,36	33
CHOR2	ERY	3,25	3,85	4,27	4,43	5,04	4,2	
KONTROLA	ERY	3,09	3,82	3,98	4,33	5,05	4,01	
CHOR1	PLT	128	179	217	266	336	225,28	
CHOR2	PLT	91	172	195	223	456	209,12	456, 314, 91, 311, 306
KONTROLA	PLT	147	188	214	254	434	225,88	434
CHOR1	HGB	9,5	11,92	12,4	13,21	14,5	12,41	0.9
CHOR2	HGB	9,83	11,76	12,57	13,69	22,23	12,81	22.23
KONTROLA	HGB	9,5	10,47	11,44	11,76	13,21	11,26	
CHOR1	HCT	0,28	0,35	0,36	0,39	0,41	0,36	0.28
CHOR2	HCT	0,04	0,33	0,36	0,39	0,41	0,35	0.04
KONTROLA	HCT	0,28	0,32	0,34	0,35	0,39	0,34	

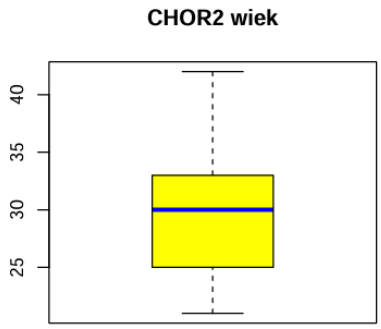
Wykresy boxplot

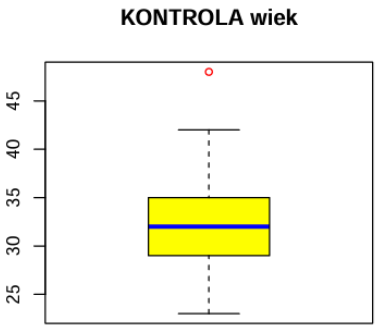
Na każdej stronie w pdf znajduję się wykresy porównawcze dla trzech grup badanych.

Odczytanie wykresu:

Na pionowej osi znajdują się wartości zmiennej. Prostokąt oznacza zakres wartosci. Dolna krawędź protokąta wyznacza pierwszy kwartyl. Górna krawędź prostokąta wyznacza trzeci kwartyl. Niebieska linia wyznacza medianę. Czerwone punkty są to wartości odstające.







Analiza różnicująca otrzymuje się dwa pliki csv. Różnice homogeniczne sa przeprowadzane przez test Turkeya po teście ANOVA.

Odczytanie wykresu:

- variable – nazwa zmiennej

- comparsion – nazwy porównywanych grup

- diff -różnica pomiędzy dwoma grupami, porównywane średnie wartości

- lwr - dolna granica przedziału ufności dla różnicy średnich

- upr - górna granica przedziału ufności dla różnicy średnich

- p adj - skorygowany poziom istotności

diff	lwr	upr	p adj
0,4232228	-0,32741093	1,173856533	0,372940393
-0,7261892	-1,47682293	0,024444533	0,060043259
-1,149412	-1,90004573	-0,39877827	0,001352322

Różnice niehomogeniczne są przeprowadzone przez test Dunna po teście Kruskala – Wallisa.

- Z - wartość statystyki
- P.unadj – nieskorygowany poziom istotności
- P.adj - skorygowany poziom istotności
- varaible –zmienna porównywana

Analiza korelacji

Opis wykresu:
Oś x to zminenna i Oś Y zmienna. P to p.value. R to wartość współczynnika korelacji. Zielona linia to linia regresji, wskazuje trend. Zielony cień opisuje zakres znalezienia lini regresji w określonym poziomie ufności. Punkty na wykresie to obserwacja danych