

Projekt R

Narzędzie służące do przeprowadzenia podstawowej analizy statystycznej dla danych medycznych z plików CSV.

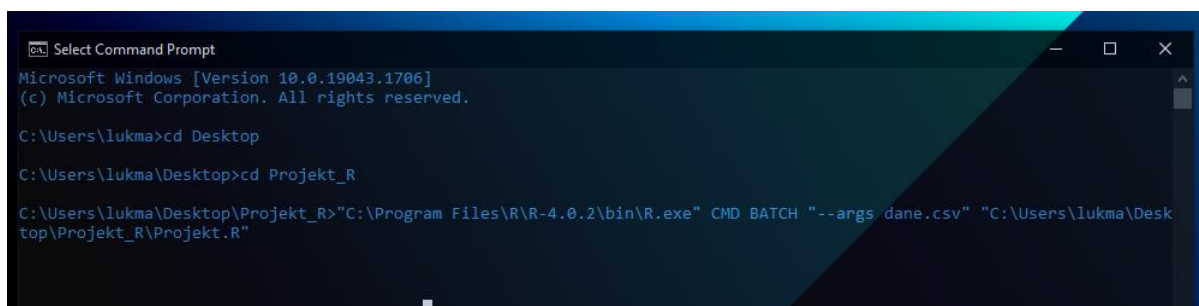
Opis Działania

Skrypt uruchamiany jest za pomocą polecenia CMD Batch Mode, które konstruujemy w następujący sposób:

"/Ścieżka_do_R.exe" CMD BATCH "--args Plik_Do_Analizy.csv" "/Ścieżka_do_skryptu_Projekt.R"

Należy jednak pamiętać, aby komenda ta została wywołana z poziomu, w którym znajduje się Plik_Do_Analizy.csv

Uwaga: batch mode nie zezwala na moim komputerze instalacje packages. Pakiety do instalacji zakomentowane.



```

Select Command Prompt
Microsoft Windows [Version 10.0.19043.1706]
(c) Microsoft Corporation. All rights reserved.

C:\Users\lukma>cd Desktop
C:\Users\lukma\Desktop>cd Projekt_R
C:\Users\lukma\Desktop\Projekt_R>"C:\Program Files\R\R-4.0.2\bin\R.exe" CMD BATCH "--args dane.csv" "C:\Users\lukma\Desktop\Projekt_R\Projekt.R"
  
```

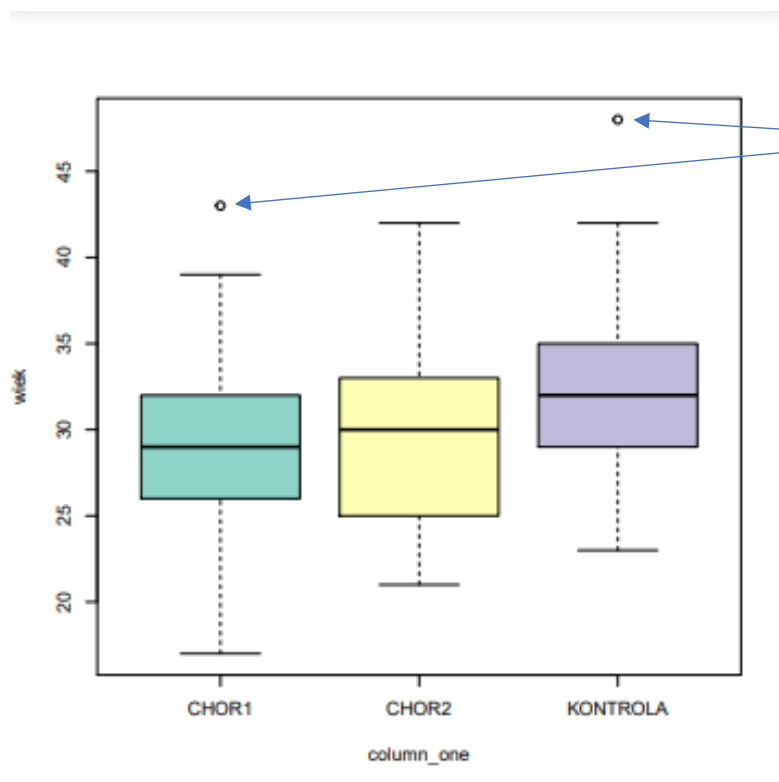
PRZYKŁADOWE WYWOŁANIE SKRYPTU.

Po wczytaniu pliku, skrypt wypełnia puste/błędne wartości poprzez wprowadzenie średniej wartości kolumny, w której błąd występuje. Wprowadzone zmiany raportowane są do pliku raport.txt, a plik final.csv zawiera dane z uzupełnionymi kolumnami.

```

#####
Made changes:
Column: 7 number of inserts: 2 value:
12.1692291780822
Column: 10 number of inserts: 1 value:
0.857027027027027
#####
  
```

Następnie do pliku outliers_pt1.pdf zapisywane zostają wykresy przedstawiające wartości odstające dla każdego parametru w przypadku każdej grupy.



Wartości odstające dla parametru „wiek” dla trzech grup z pliku dane.csv

Natomiast do pliku groups_characteristics_pt2.pdf zapisywane są ogólne charakterystyczne cechy dla każdego parametru. Na poniższym przykładzie dla hsCRP wypisane zostają ilość przypadków („count”), średnia wartości dla grupy („mean”), odchylenie standardowe („sd”), mediana („median”), oraz wyniku testu Shapiro-Wilk, który zostanie jeszcze zaprezentowany w kolejnych krokach.

hsCRP

	grupa	count	mean	sd	median	Shapiro.stat	Shapiro.p_value
1	CHOR1	25	6.10	8.82	3.97	0.5418468	9.933661e-08
2	CHOR2	25	5.54	4.65	3.45	0.8731146	4.998515e-03
3	KONTROLA	25	5.30	4.00	4.22	0.8813002	7.351605e-03

Kolejne zostają wykonane analizy porównawcze. Rodzaje testów zależne są od ilości grup w analizowanym pliku, oraz czy dane nadają się do zastosowania testów parametrycznych czy nie.

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 (<i>post hoc</i> Tukeya)
		NIE	test Kruskala-Wallisa (<i>post hoc</i> Dunna)
	NIE	-	

SCHEMAT WYBORU TESTÓW STATYSTYCZNYCH

W pierwszym kroku sprawdzana jest ilość niezależnych grup, a następnie zgodność z rozkładem normalnym badanego parametru. Hipotezy testów normalności rozkładu:

-H0 : rozkład badanej cechy w populacji jest rozkładem normalnym

-H1 : rozkład badanej cechy w populacji jest różny od rozkładu normalnego

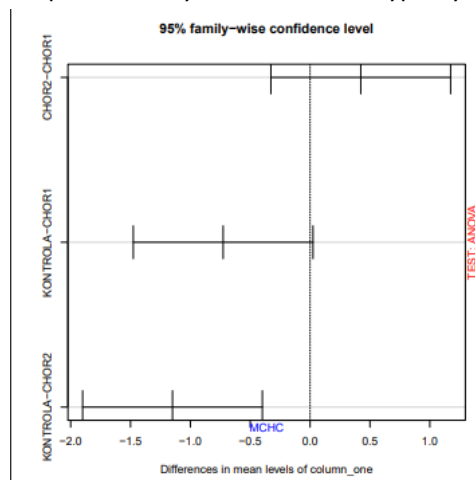
Wyznaczoną na podstawie testu Shapiro-Wilka wartość p porównujemy z poziomem istotności α :

- jeżeli $p \leq \alpha$ to odrzucamy H0, przyjmując H1

- jeżeli $p > \alpha$ nie ma podstaw aby odrzucić H0

Kolejne kryterium to jednorodność wariancji, inaczej homogeniczność wariancji. Heterogeniczność wariancji oznacza, że wariancje populacji porównywanych grup lub komórek nie są jednorodne ani równe. Aby zweryfikować, czy grupy mają różne wariancje skrypt wykonuje test Levene'a . Jeżeli wynik testu Levene'a jest istotny statystycznie ($p < 0,05$) oznacza to, że wariancje nie są podobne (są heterogeniczne).

Na podstawie tych warunków skrypt wykonuje odpowiednie testy parametryczne i nieparametryczne.



Przykładowy Test parametryczny
ANOVA dla MCHC w pliku
groups_comparison_pt3.pdf

Wyniki dla każdego parametru są kolejno zapisywane w pliku raport.txt

W ostatnim etapie dane poddane są testom badającym korelację. Ze względu na zgodność z rozkładem normalnym test korelacji może odbyć się za pomocą metody Pearsona (Shapiro $p.value > 0.05$), jak i Spearmana (Shapiro $p.value < 0.05$). Uzyskana wartość p testu korelacji świadczy o obecności lub braku korelacji, natomiast funkcja `correlation<-function(x)`, przyjmując w argumencie współczynnik korelacji r zwraca interpretację siły korelacji.