Rozdział 5 - Analiza statystyczna - ćwiczenia

## Analiza statystyczna opisowa pliku *03 - analiza płac.csv*

1. Wczytanie i transformacja pliku csv *03 - analiza płac.csv*.

rm(list=ls())  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
Dane<-read.table("./Dane/03 - analiza płac.csv",header = T,  
 dec=",",sep = ";",encoding = "UTF-8")  
colnames(Dane)[1]<-"Płeć"  
colnames(Dane)[5]<-"Płaca"  
tbl\_df(head(Dane))

## # A tibble: 6 x 5  
## Płeć Wykształcenie Wiek Staż.pracy Płaca  
## <fct> <fct> <int> <int> <dbl>  
## 1 Kobieta Podstawowe 42 NA 420   
## 2 Kobieta Wyższe 34 2 1268.  
## 3 Kobieta Średnie zawodowe NA 17 862   
## 4 Kobieta Zasadnicze 44 25 662.  
## 5 Kobieta Średnie zawodowe 38 5 543.  
## 6 Kobieta Wyższe 45 NA 816

1. Statystyka opisowa.

Dane %>% group\_by(Wykształcenie,Płeć) %>%   
 summarize(  
 temp=list(c(Count=n(),  
 round(summary(Płaca),2),  
 "St dev"=round(sd(Płaca),2)))) %>%   
 unnest\_wider(temp)

## # A tibble: 10 x 10  
## # Groups: Wykształcenie [5]  
## Wykształcenie Płeć Count Min. `1st Qu.` Median Mean `3rd Qu.` Max.  
## <fct> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Podstawowe Kobi~ 31 288. 420 466 484. 519 1250   
## 2 Podstawowe Mężc~ 50 239 592. 650. 650. 741. 957.  
## 3 Średnie ogól~ Kobi~ 96 300 602. 702 748. 850 2226   
## 4 Średnie ogól~ Mężc~ 14 433. 824. 878. 962. 1125 1854   
## 5 Średnie zawo~ Kobi~ 415 250 548. 680 737. 850 2355   
## 6 Średnie zawo~ Mężc~ 156 210 640 777 881. 1042. 2300   
## 7 Wyższe Kobi~ 151 400 708. 900 970. 1165 3000   
## 8 Wyższe Mężc~ 128 276 1000 1200 1232. 1421. 2767.  
## 9 Zasadnicze Kobi~ 72 150 355 458. 516. 652. 1130   
## 10 Zasadnicze Mężc~ 142 260 501 595. 687. 810. 2400   
## # ... with 1 more variable: `St dev` <dbl>

## Weryfikacja hipotez statystycznych pliku *03 - analiza płac.csv*

1. Test na normalność rozkładów płacy dla kobiet i mężczyzn.

wynik<-Dane %>% group\_by(Płeć) %>%   
 summarize(  
 temp=list(c(round(as.numeric(shapiro.test(Płaca)[1:2]),4),  
 Rozkład=ifelse(shapiro.test(Płaca)[2]>0.05,"Normalny","Inny")))) %>% unnest\_wider(temp)

## New names:  
## \* `` -> ...1  
## \* `` -> ...2  
## New names:  
## \* `` -> ...1  
## \* `` -> ...2

colnames(wynik)<-c("Płeć","Statystyka S-W","p-value","Rozkład")  
wynik

## # A tibble: 2 x 4  
## Płeć `Statystyka S-W` `p-value` Rozkład  
## <fct> <chr> <chr> <chr>   
## 1 Kobieta 0.8666 0 Inny   
## 2 Mężczyzna 0.909 0 Inny

Czyli dalej analizujemy to testami nieparametrycznymi pod kątem tego czynika.

1. Test na równość median dla populacji kobiet i mężczyzn.

data.frame(wilcox.test(Dane$Płaca~Dane$Płeć)[c(1,3)],  
 "Mediany"=ifelse(wilcox.test(Dane$Płaca~Dane$Płeć)[3]>0.05,"Równe","Nierówne"))

## statistic p.value Mediany  
## W 149004 8.576489e-10 Nierówne

Czyli **płeć** jest czynnikiem mającym wpływ na pesję dlatego, że **p-value**<0.05. Mediany pensji w populacjach kobiet i mężczyzn nie są sobie równe - różnią się w sposób istotny statystycznie.

1. Test na proporcjonalność rozkładu płeć vs wykształcenie.

temp<-table(Dane$Wykształcenie,Dane$Płeć) #tabela krzyżowa  
chisq.test(temp)

##   
## Pearson's Chi-squared test  
##   
## data: temp  
## X-squared = 155.04, df = 4, p-value < 2.2e-16

chisq.test(temp)$expected #krzyżowa tabela oczekiwana

##   
## Kobieta Mężczyzna  
## Podstawowe 49.37450 31.62550  
## Średnie ogólne 67.05179 42.94821  
## Średnie zawodowe 348.05976 222.94024  
## Wyższe 170.06773 108.93227  
## Zasadnicze 130.44622 83.55378

chisq.test(temp)$observed #krzyżowa tabela wejściowa

##   
## Kobieta Mężczyzna  
## Podstawowe 31 50  
## Średnie ogólne 96 14  
## Średnie zawodowe 415 156  
## Wyższe 151 128  
## Zasadnicze 72 142

Ten test dla zmiennych jakościowych wykazał, że bazując na **p-value**<0.05, proporcjonalność rozkładu cechy wykszałcenie dla kobiet i mężczyzn nie jest taka sama.

1. Test na rówość median dla wielu populacji. Czy poziom wykształcenia ma wpływ na pensje?
2. Test na normalność rozkładów płac z podziałem na wykształcenie.

wynik<-Dane %>% group\_by(Wykształcenie) %>%   
 summarize(  
 temp=list(c(round(as.numeric(shapiro.test(Płaca)[1:2]),4),  
 Rozkład=ifelse(shapiro.test(Płaca)[2]>0.05,"Normalny","Inny")))) %>% unnest\_wider(temp)

## New names:  
## \* `` -> ...1  
## \* `` -> ...2  
## New names:  
## \* `` -> ...1  
## \* `` -> ...2  
## New names:  
## \* `` -> ...1  
## \* `` -> ...2  
## New names:  
## \* `` -> ...1  
## \* `` -> ...2  
## New names:  
## \* `` -> ...1  
## \* `` -> ...2

colnames(wynik)<-c("Wykształcenie","Statystyka S-W","p-value","Rozkład")  
wynik

## # A tibble: 5 x 4  
## Wykształcenie `Statystyka S-W` `p-value` Rozkład  
## <fct> <chr> <chr> <chr>   
## 1 Podstawowe 0.9674 0.037 Inny   
## 2 Średnie ogólne 0.8559 0 Inny   
## 3 Średnie zawodowe 0.8659 0 Inny   
## 4 Wyższe 0.9254 0 Inny   
## 5 Zasadnicze 0.8682 0 Inny

Czyli dalej analizujemy to testami nieparametrycznymi pod kątem tego czynika.

1. Test na rówość median dla wielu populacji.

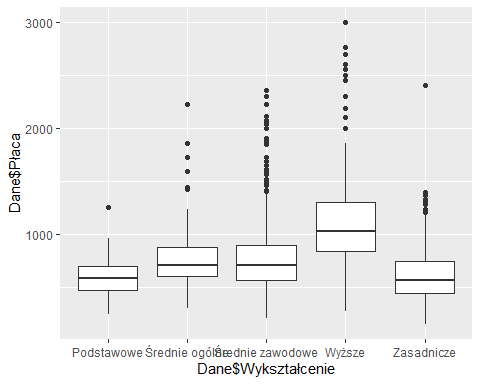
kruskal.test(Dane$Płaca~Dane$Wykształcenie)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Dane$Płaca by Dane$Wykształcenie  
## Kruskal-Wallis chi-squared = 257.56, df = 4, p-value < 2.2e-16

pairwise.wilcox.test(Dane$Płaca,Dane$Wykształcenie,p.adj="bonf")

##   
## Pairwise comparisons using Wilcoxon rank sum test   
##   
## data: Dane$Płaca and Dane$Wykształcenie   
##   
## Podstawowe Średnie ogólne Średnie zawodowe Wyższe   
## Średnie ogólne 7.1e-06 - - -   
## Średnie zawodowe 2.1e-07 1 - -   
## Wyższe < 2e-16 3.0e-13 < 2e-16 -   
## Zasadnicze 1 4.5e-06 1.2e-10 < 2e-16  
##   
## P value adjustment method: bonferroni

library(ggplot2)  
ggplot(Dane,aes(y=Dane$Płaca,x=Dane$Wykształcenie))+geom\_boxplot()



Nie różnią się tylko populacje wykształcenia na poziomie ***zasadniczym*** z ***podstawowym*** oraz ***średnim*** z ***średnim zawodowym***.

## Estymacja przedziałowa pliku *03 - analiza płac.csv*

Ile wynosi średnia pensja w każdej populacji różnych poziomów wyksztalcenia?

Dane %>% group\_by(Wykształcenie) %>% mutate(Ufność=qt(1-0.05/2,n()-1)\*  
 sd(Płaca)/sqrt(n()-1)) %>%  
 summarise(Dolna=mean(Płaca)-min(Ufność),Średnia=mean(Płaca),  
 Górna=mean(Płaca)+min(Ufność))

## # A tibble: 5 x 4  
## Wykształcenie Dolna Średnia Górna  
## <fct> <dbl> <dbl> <dbl>  
## 1 Podstawowe 549. 586. 624.  
## 2 Średnie ogólne 717. 775. 834.  
## 3 Średnie zawodowe 749. 776. 803.  
## 4 Wyższe 1038. 1090. 1142.  
## 5 Zasadnicze 592. 630. 667.

## Ćwiczenia

1. Dla zestawu danych 02 - dane.csv (wygenerowanych w poprzednich rozdziałach) wykorzystać znane miary statystyk opisowych do raportu ilości przepracowanych godzin z podziałem na przedziały wie-kowe, wykształcenie, płeć i rasę.
2. Raport z poprzedniego zadania podeprzeć histogramami i wykresami pudełkowymi,
3. Dla danych z poprzednich zadań pokazać liczbowo oraz na wykresie, czy jest jakaś zależność między wiekiem a ilością przepracowanych godzin w tygodniu.
4. Dla danych z poprzedniego zadania wykonać estymację przedziałową średniego wieku kobiet i męż-czyzn dodając także podział na poziom wykształcenia i rasę.
5. Wczytaj plik 03 - kredyty.xls.
6. Przekoduj zmienne czynnikowe (Limit\_na\_karcie, Plec (1 = mężczyzna; 2 = kobieta), Stan\_cywilny (1 = w\_zwiazku; 2 = kawaler\_panna; 3 = inny), Wykształcenie (1 = podstawowe; 2 = wyższe; 3 = śred-nie; 4 = inne), Wiek (podany w latach).
7. Usuń wartości odstające w każdej grupie (Płeć/Wykształcenie/Stan\_cywilny).
8. Sprawdź, czy dane mają rozkład normalny.
9. Zbadaj, czy czynniki Płeć/Wykształcenie/Stan\_cywilny mają wpływ na limit na karcie,
10. Jaki ma wpływ usuwanie skośności lub standaryzacja na wnioski z powyższej analizy.