Mini raport analityczny

Zofia Samsel*

14 grudzień, 2022

Analiza wyników sondarzu społecznego European Social Servey z roku 2018.

Branie udziału w wyborach a stopień przekonania o wpływie jednostki na system polityczny

Czy występuje zależność między zmienną porządkową pspipla a zmienną nominalną vote01?

Opis zmiennych:

- zmienna porządkowa **psppipla** - stopień przekanania o wpływie na system polityczny (Not at all; Very little; Some; A lot; A great deal)
- zmienna nominalna **vote01** - czy osoba głosowała? (YES, NO)

W pierwszym kroku ładuję dane i prezentuje je w tabeli:

```
#przygotowywanie danych
df$vote01 = df$vote
df$vote01[df$vote01 == "Not eligible to vote"] = NA

#usunięcie NA z danych vote01
df$vote01 = droplevels(df$vote01)

#usuwanie NA z dancyh psppipla01
df$psppipla01 = droplevels(df$psppipla)


#prezentacja w tabeli
tbl1 = table(df$psppipla01,df$vote01)
tbl1
```

```
##
##                Yes  No
##   Not at all   202 120
##   Very little  382 143
##   Some         293  82
##   A lot         57  13
##   A great deal  12   2
```

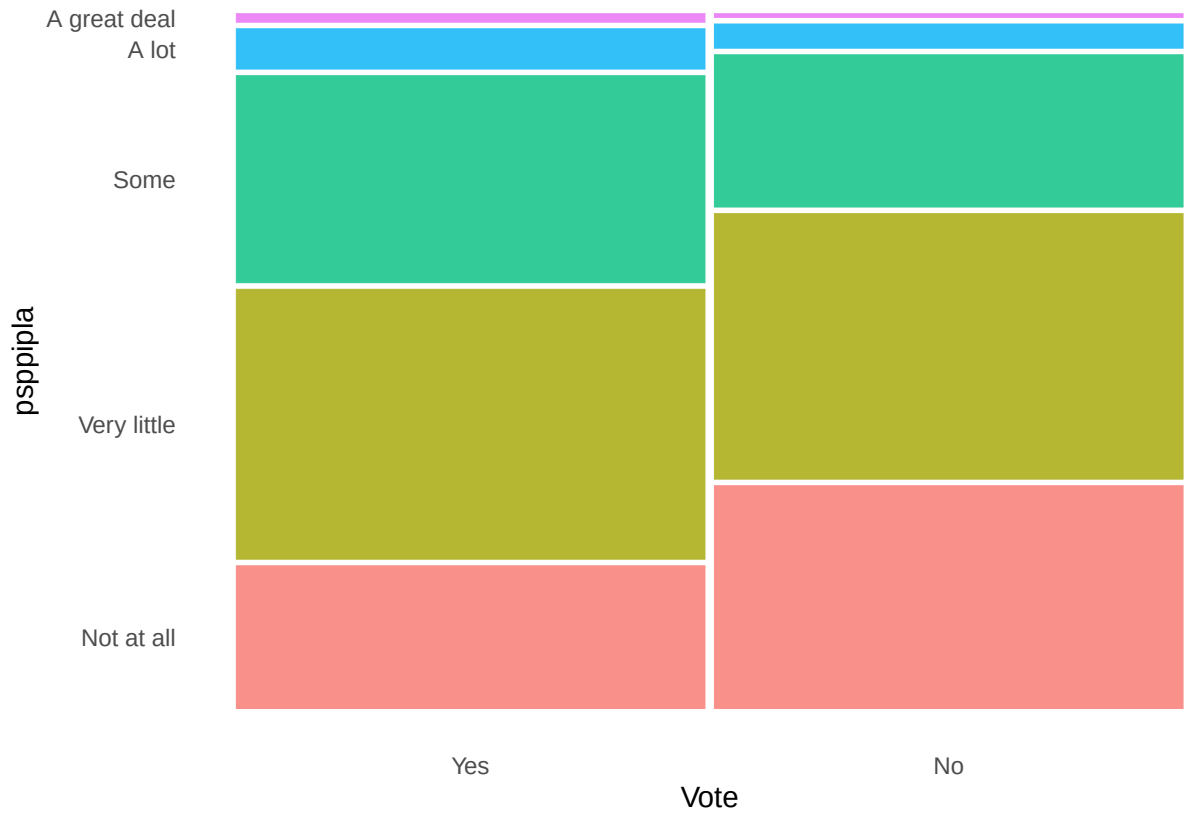
Dla lepszego przeanalizowania danych generuję wykres mozaikowy przedstawiający jak rozkładają się tendencje do głosowania pod wzgledem przekonania o tym, czy jednostka ma wpływ na system polityczny.

```
#tworzenie dataframe z danymi
tbl_df = df %>%
  filter(!is.na(vote01), !is.na(psppipla01))
```

*zofia.samsel@student.uj.edu.pl

```
#rysowanie wykresu
ggplot(data = tbl_df) +
  geom_mosaic(aes(x=product(psppipla01), fill = psppipla01,
                  conds = product(vote01))) +
  labs(x = 'Vote', y = 'psppipla') +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank()) +
  theme(axis.ticks.y=element_blank(), axis.ticks.x=element_blank()) +
  theme(legend.position = "none")
```

```
## Warning: `unite_()` was deprecated in tidyr 1.2.0.
## i Please use `unite()` instead.
## i The deprecated feature was likely used in the ggmosaic package.
##   Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
```



Z wykresu można odczytać pewne tendencje. Osoby, które nie biorą udziału częściej odpowiadały, że nie wierzą w swój wpływ na system polityczny. Z drugiej strony było więcej osób, które brały udział w wyborach i jednocześnie uważały, że mają wpływ na system w porównaniu do osób, które nie uczestniczyły w wyborach.

Sprawdzam, czy wyżej opisane różnice są znaczące statystycznie.

```
#tworzenie tabeli
tbl1 = table(df$psppipla01,df$vote01)
#chi2 test
chisq.test(tbl1)
```

```
## Warning in chisq.test(tbl1): Aproksymacja chi-kwadrat może być niepoprawna
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl1
## X-squared = 25.379, df = 4, p-value = 4.221e-05
```

```
chisq.test(tbl1,filter,sim=T,B=1000) #bootstrapped chi2
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 1000
##  replicates)
##
## data:  tbl1
## X-squared = 25.379, df = NA, p-value = 0.000999
```

```
#Fisher test
fisher = fisher.test(tbl1, simulate.p.value=TRUE)
fisher #OR=1 oznacza brak związku
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  2000 replicates)
##
## data:  tbl1
## p-value = 0.0004998
## alternative hypothesis: two.sided
```

Wartość krytyczna testu t dla danych charakteryzujących się 4 stopniami swobody i alfą 0,05 wyniosi około 9,5. Wartość testu Chi-kwadrat wynosi 25.4 co pozwala na odrzucienie hipoteyę zerowej o braku zależności między danymi (p-value < 0.001). Test chi-kwadrat wykazał zatem, że istnieje istotna zależność między przekonaniem o wypływnie na system polityczny a braniem udziału w wyborach.

Test chi- kwadrat z symulowanym p-value na podstawie 1000 powtórzeń był równeiż większy niż 9.5 i wynosił 25,4 (p-value < 0.001). Pozwala to na odrzucie hipotezy zerowej i potwierdzenie zależności między danymi.

Test Fishera wykazał istotną zależność między zmiennymi (p-value $= 0.0005$).

Powyższe testy potwierdzają istnienie istotnej zależność między przekonaniem o wypływnie na system polityczny a faktem, czy osoba bierze udziału w wyborach.

Branie udziału w wyborach a poziom zaufania względem policji

Czy występuje zależność między ufnością policji a braniem udziału w wyborach?

Opis zmiennych:

- zmienna niezależna **votes**: zmienna nominalna vote01 - czy osoba bierze udział w wyborach
- zmienna zależna **police**: zmienna ilościowa trstplc - jak bardzo osoba ufa policji (NO 0-10 complete trust)

W pierwszym kroku przygotowuje zmienne i analizuje ich statystyki opisowe:

```
#przygotowanie zmiennych

#usuwanie braków w trstplc
levels(df$trstplc)
```

```
##  [1] "No trust at all" "1"             "2"             "3"
##  [5] "4"               "5"             "6"             "7"
```

```
##  [9] "8"                "9"                "Complete trust"
```

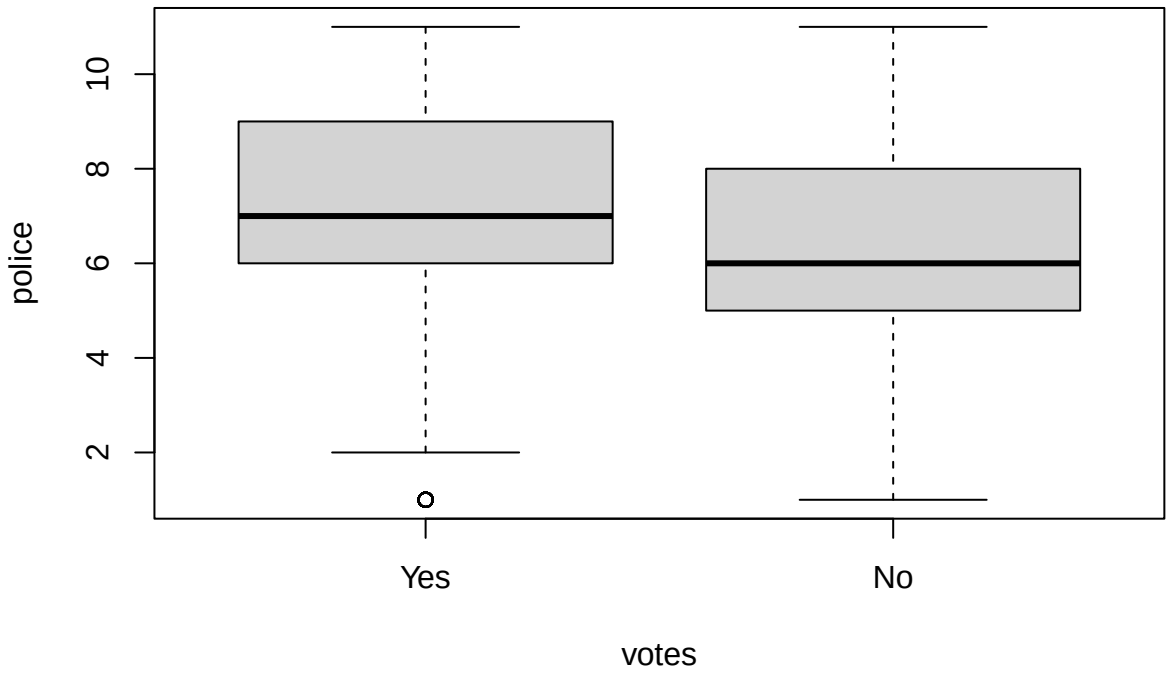
```
#tworzenie zmiennych
police = as.numeric(df$trstplc)
votes = df$vote01

#ładowanie statysyk
describeBy(police, group=votes)
```

```
##
##  Descriptive statistics by group
## group: Yes
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 951 6.91 2.25      7    7.03 1.48   1  11    10 -0.52     0.21 0.07
## ------------------------------------------------------------
## group: No
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 375 6.41 2.42      6    6.49 2.97   1  11    10 -0.29    -0.36 0.12
```

Dla lepszego przenalizowania danych tworzę wykres boksowy.

```
#wykres boksowy
boxplot(police ~ votes)
```



Z wykresu można zaobserwować, że istnieje różnica między grupą, która wzięła udział w wyborach a tą, która nie głosowała pod względem poziomu zaufania do policji.

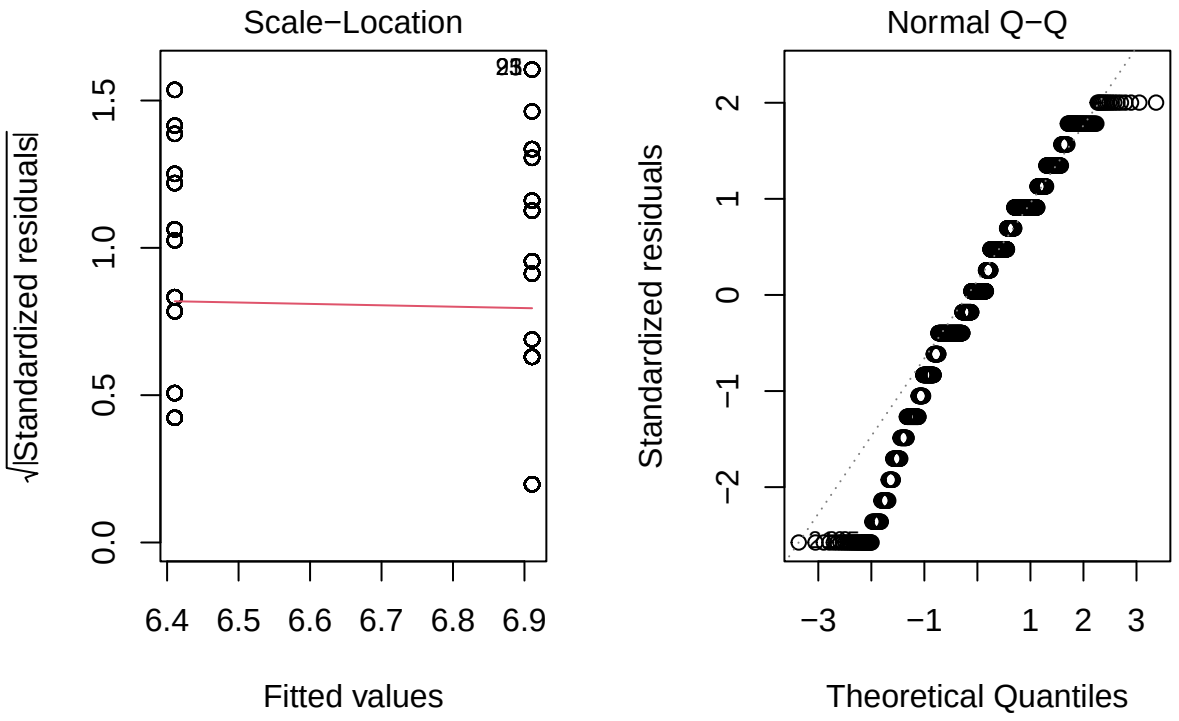
W celu sprawdzenia istotności tej różnicy przeprowadzę testy statystyczne. W pierwszym kroku sprawdzam, czy dane układają się w rozkład normalny.

```r
#test rozkładu normanego
shapiro.test(police) #Shapiro-Wilk normality test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  police
## W = 0.95788, p-value < 2.2e-16
```

Można to lepiej zobaczyć na wykresach:

```r
#Wykresy
res_aov = aov(police ~ votes) #ANOVA
par(mfrow = c(1, 2)) # combine plots
plot(res_aov, which = 3) # 1. Homogeneity of variances
plot(res_aov, which = 2) # 2. Normality
```



Ponieważ dane rozkładaja się prawie normalnie możemy przeporwadzić test te do sprawdzenia istnienia zależnosci miedzy zmiennymi police ~ votes.

```r
# test t
t.test(police ~ votes)
```

```
##
##  Welch Two Sample t-test
##
## data:  police by votes
## t = 3.4586, df = 642.44, p-value = 0.0005788
## alternative hypothesis: true difference in means between group Yes and group No is not equal to 0
```

```
## 95 percent confidence interval:
##  0.2161021 0.7838054
## sample estimates:
## mean in group Yes  mean in group No
##          6.910620          6.410667
```

```
#ANOVA
report(res_aov)
```

```
## Warning: Could not find Sum-of-Squares for the (Intercept) in the ANOVA table.
```

```
## The ANOVA (formula: police ~ votes) suggests that:
##
##   - The main effect of votes is statistically significant and very small (F(1,
## 1324) = 12.75, p < .001; Eta2 = 9.54e-03, 95% CI [2.78e-03, 1.00])
##
## Effect sizes were labelled following Field's (2013) recommendations.
```

```
# test Shapiro-Wilk dla ANOVA
shapiro.test(res_aov$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov$residuals
## W = 0.97175, p-value = 1.966e-15
```

Test t wykazał, że istnieje istotna zależność miedzy faktem, czy ktoś głosuje, a jego poziomem zaufania do policji (95% Cl [0.2161021, 0.7838054], p = 0.0005788).

Dodatkowo, ANOVA potwierdziłą, że efekt główny zmiennej niezależnej jest istotny i mały ((F(1, 1324) = 12.75, p < .001; Eta2 = 9.54e-03, 95% CI [2.78e-03, 1.00])).

Podsumowując, istnieje zależość miedzy zaufaniem do policji a faktem, czy ktoś brał udział w wyborach.

Czas korzystania z telefonu a przekonaie o tym, czy ludzie są bardziej pomocni, czy samolubni

Czy istnieje zależność miedzy czasem korzystania z telefonu a przekonamiem, że ludzie są bardziej pomocni lub bardziej samolubni?

Opis zmiennych:

- zmienna niezależna **komp**: zmienna ilościowa netustm - jak dużo osoba korzysta z komputera (w minutach)
- zmienna zależna **ufnosc**: zmienna ilościowa pplhlp - jak bardzo uważasz, że lubie starają się być pomocni, czy raczej martwią się tylko o siebie? (ludzie dbają tylko o siebie 0-10 ludzie są pomocni)

W pierwszym kroku tworzę dataframe z danymi oraz analizuje statystyki opisowe:

```
#tworzenie tabeli danych
reg_df = df[, c("netustm", "pplhlp")]
names(reg_df)[1:2] = c("komp", "ufnosc")

#modyfikowanie zmiennych na numeryczne
reg_df = as.data.frame(sapply(reg_df, as.numeric))

#wyświetlanie statystyki opisowej
summary(reg_df)
```

```
##       komp           ufnosc
##  Min.   : 1.00   Min.   : 1.00
##  1st Qu.:12.00   1st Qu.: 3.00
##  Median :16.00   Median : 5.00
##  Mean   :20.69   Mean   : 4.98
##  3rd Qu.:30.00   3rd Qu.: 6.00
##  Max.   :44.00   Max.   :11.00
##  NA's   :647     NA's   :8
```

Tworzę model regresji liniowej ufnosc ~ komp oraz prezentuje informacje o modelu.

```
#tworzenie modelu
model1 = lm(ufnosc ~ komp, data=reg_df)
#model1 #Print the regression model
summary(model1)
```

```
##
## Call:
## lm(formula = ufnosc ~ komp, data = reg_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5507 -1.5319  0.0879  1.5245  6.2005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.743132   0.168180  28.203   <2e-16 ***
## komp        0.018780   0.007204   2.607   0.0093 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.274 on 849 degrees of freedom
##   (649 obserwacji zostało skasowanych z uwagi na braki w nich zawarte)
## Multiple R-squared:  0.007941,   Adjusted R-squared:  0.006772
## F-statistic: 6.796 on 1 and 849 DF,  p-value: 0.009299
```

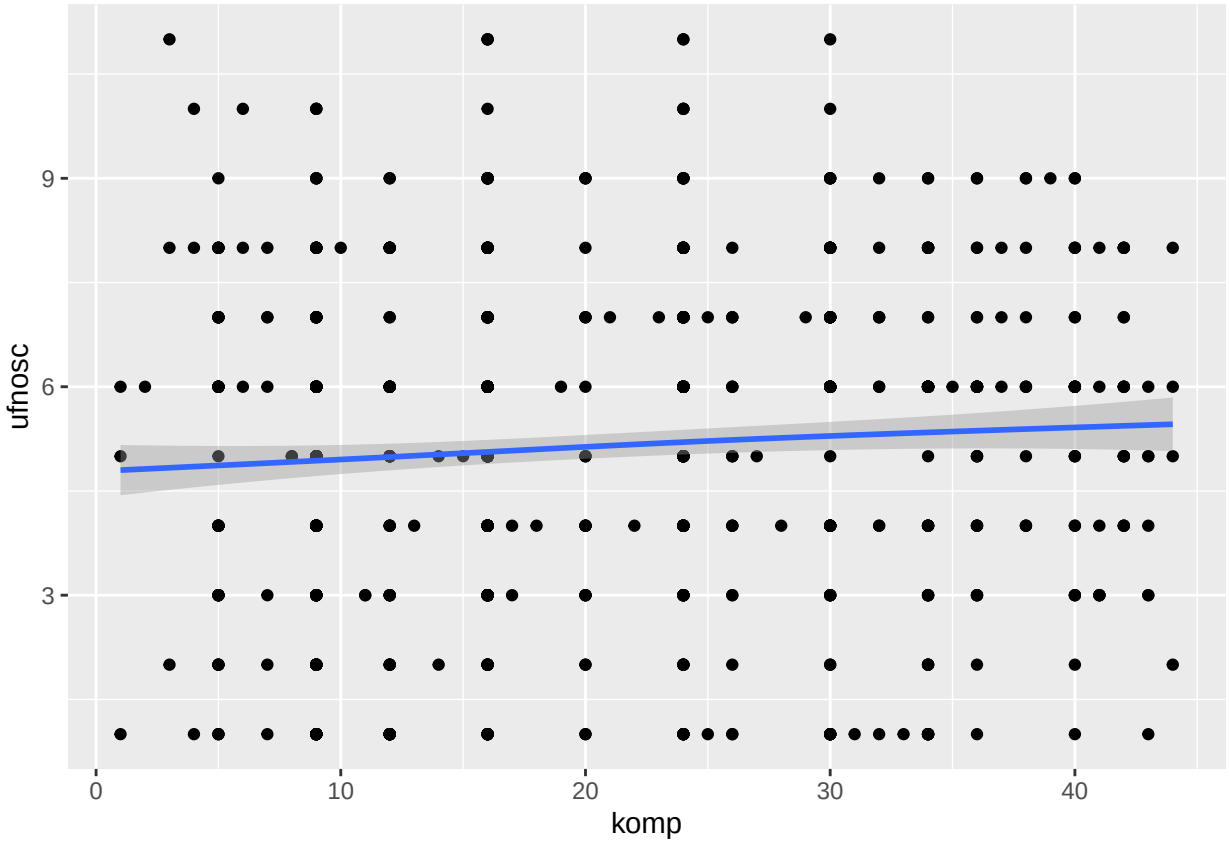
Żeby lepiej przeanalizować dane tworzę wykres modelu ufnosc ~ komp:

```
#wykres danych
ggplot(reg_df, aes(x = komp, y = ufnosc)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 649 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 649 rows containing missing values (`geom_point()`).
```



Dane są rozłożone w mniej więcej równomierną chmurę. Linia regresji nie wyglada też jakby miała wystapić zależność między danymi. Dla pewności sprawdzam, czy istnieje korelacja między czasem korzystania z telefonu a przekoaniem co do pomocy innym.

```
#liczenie korelacji
cor(reg_df, use = "pairwise.complete.obs")
```

```
##              komp     ufnosc
## komp   1.00000000 0.08910992
## ufnosc 0.08910992 1.00000000
```

Test korelacji nie wykrył istotnego związku liniowego między czasem korzystania z telefonu a przekonamiem, że ludzie są bardziej pomocni lub bardziej samolubni (r = 0.089).