

# PODSUMOWANIE

*Szymon Zdanowski*

*21 marca 2018*

## W SKRÓCIE

W ankiecie wzięły udział **792 osoby**. W tym 416 studentów kierunków medycznych i 376 studentów kierunków niemedycznych. 74.87% (593) respondentów było kobietami.

Analizy zmiennych ilościowych dokonano za pomocą testu U Manna-Whitneya oraz testu Chi Kwadrat. Analizy zmiennych binarnych dokonano za pomocą testu Chi Kwadrat. Przyjęty poziom istotności - alfa = 0,05. Szczegółowy opis zmiennych można znaleźć pod poniższym linkiem.

## >>SZCZEGÓŁOWY OPIS ZMIENNYCH<<

### 1. Zmienne ilościowe

Użyto testu U manna-whitneya (wilcoxon), ponieważ zmienne te nie mają rozkładu normalnego (ciężko żeby miały przy 5 wartościach). Między osobami sięgającymi po substancje nielegalne a niesięgającymiu dokonano porównania w odpowiedzi na pytania:

W jakim stopniu zgadzasz się ze stwierdzeniem: "Ktoś usilnie namawiał mnie do pójścia na studia/wybrania tego kierunku." = mienna **Namawianie**

W jakim stopniu zgadzasz się ze stwierdzeniem: "Moje studia są ciekawe i interesujące."? = zmienna **Ciekawe**  
Przypisano odpowiednie zakresy danych.

legal = grupa nieużywająca substancji nielegalnych podczas nauki

nielegal = grupa używająca substancji nielegalnych podczas nauki

```
load(file = "dane.Rda")
legal <-
  subset.data.frame(dane, subset = dane$nielegal_ogol == 0)
#subset NIEsiegajacych po nielegalne substancje
nielegal <-
  subset.data.frame(dane, subset = dane$nielegal_ogol == 1)
#subset siegajacych po nielegalne substancje
```

### Pytanie pierwsze

H0 - nie ma istonej różnicy między grupami H1 - istnieje statystycznej istotna różnica między grupami

```
wilcox.test(legal$Namawianie, nielegal$Namawianie)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: legal$Namawianie and nielegal$Namawianie
## W = 17010, p-value = 0.01582
## alternative hypothesis: true location shift is not equal to 0
```

p-value = 0.01582,  $p > 0.05$   $H_0$  odrzucona. Zatem istnieje istotna statystycznie różnica między grupami. Średnia odpowiedź w grupie nielegal

```
## [1] "2.34 ± 1.52"
```

Średnia odpowiedź w grupie legal

```
## [1] "1.83 ± 1.17"
```

Zatem grupa nielegal uzyskała istotnie wyższy wynik. **Wiem, że porównuje zmienne o rozkładzie innym niż normalny ale co mam zrobić jeżeli mediany są takie same?**

### Pytanie drugie

$H_0$  - nie ma istotnej różnicy między grupami  $H_1$  - istnieje statystycznie istotna różnica między grupami

```
wilcox.test(legal$Ciekawe, nielegal$Ciekawe)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  legal$Ciekawe and nielegal$Ciekawe
## W = 23814, p-value = 0.03942
## alternative hypothesis: true location shift is not equal to 0
```

p-value = 0.03942,  $p > 0.05$   $H_0$  odrzucona. Zatem istnieje istotna statystycznie różnica między grupami. Średnia odpowiedź w grupie nielegal

```
## [1] "3.68 ± 1.24"
```

Średnia odpowiedź w grupie legal

```
## [1] "4.06 ± 0.96"
```

**Wiem, że porównuje zmienne o rozkładzie innym niż normalny ale co mam zrobić jeżeli mediany są takie same?**

### Porównanie wieku między grupami

```
wilcox.test(legal$Wiek, nielegal$Wiek)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  legal$Wiek and nielegal$Wiek
## W = 20240, p-value = 0.8214
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(legal$rok, nielegal$rok)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  legal$rok and nielegal$rok
## W = 22659, p-value = 0.2033
## alternative hypothesis: true location shift is not equal to 0
```

W obydwu przypadkach  $p > 0.5$ , ZATEM brak różnic pod tym względem.

Porównanie rozkładu płci:

Analizuję tu zależność między zmiennymi różnymi zmiennymi binarnymi (Oprócz "Wiek" "Namawianie" i "Ciekawe" które są w skali likerta) a zmienną "nielegal\_ogol". >>SZCZEGÓŁOWY OPIS ZMIENNYCH<<

“In `chisq.test(tab)` : Chi-squared approximation may be incorrect”?

## Przypomnienie numeracji kolumn w dane.Rda

[illegible]

```

istotnenumer <-
  c(istotnenumer, i)
#Tworzy wektor z numerami kolumn(zmiennych) dla których p < 0.5
istotnenazwa <-
  c(istotnenazwa, names(dane[i]))
#Tworzy wektor z nazwami kolumn(zmiennych) dla których p < 0.5
}
print(
  "
  ,
  quote = FALSE
) #może da się to inaczej zrobić?
}

```

```

## [1] "2 - Wiek  a nielegal_ogol"
##
## Pearson's Chi-squared test
##
## data:  tab
## X-squared = 32.819, df = 18, p-value = 0.01755
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "3 - M=0/K=1  a nielegal_ogol"
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 7.2575, df = 1, p-value = 0.007061
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "6 - rok  a nielegal_ogol"
##
## Pearson's Chi-squared test
##
## data:  tab
## X-squared = 5.8335, df = 5, p-value = 0.3228
##
## [1] -----
## [1] "7 - Namawianie  a nielegal_ogol"
##
## Pearson's Chi-squared test
##
## data:  tab
## X-squared = 14.827, df = 4, p-value = 0.005075
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "8 - Ciekawe  a nielegal_ogol"
##
## Pearson's Chi-squared test
##
## data:  tab

```

```

## X-squared = 14.187, df = 4, p-value = 0.006722
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "10 - Medycyna a nielegal_ogol"
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: tab
## X-squared = 1.1871, df = 1, p-value = 0.2759
##
## [1] -----
## [1] "11 - Kierunek_medyczny a nielegal_ogol"
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: tab
## X-squared = 6.1233, df = 1, p-value = 0.01334
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "12 - Kofeina a nielegal_ogol"
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: tab
## X-squared = 7.0656, df = 1, p-value = 0.007858
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "13 - Nikotyna a nielegal_ogol"
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: tab
## X-squared = 74.794, df = 1, p-value < 2.2e-16
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "14 - alkohol a nielegal_ogol"
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: tab
## X-squared = 69.001, df = 1, p-value < 2.2e-16
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "15 - napoje_energetyzujące a nielegal_ogol"
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: tab
## X-squared = 10.009, df = 1, p-value = 0.001557

```

```
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
## [1] "16 - dopalacze a nielegal_ogol"
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: tab
## X-squared = 67.678, df = 1, p-value < 2.2e-16
##
## [1] p < 0.5 ISTOTNOŚĆ!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
## [1] -----
options(warn=0)
```

Zmienne (i ich numery kolumn) dla których w teście chisquare  $p < 0.05$

```
## function (x, as.factor = FALSE)
## {
##   if (as.factor) {
##     labs <- colnames(x, do.NULL = FALSE, prefix = "")
##     res <- factor(.Internal(col(dim(x))), labels = labs)
##     dim(res) <- dim(x)
##     res
##   }
##   else .Internal(col(dim(x)))
## }
## <bytecode: 0x000000001189d630>
## <environment: namespace:base>

##      [,1] [,2]      [,3]      [,4]      [,5]
## [1,] "Wiek" "M=0/K=1" "Namawianie" "Ciekawe" "Kierunek_medyczny"
## [2,] " 2"   " 3"       " 7"         " 8"      "11"

##      [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] "Kofeina" "Nikotyna" "alkohol" "napoje_energetyzujące" "dopalacze"
## [2,] "12"      "13"      "14"      "15"      "16"
```

### 3. Regresja logistyczna

Za radą Dr. Bandurskiego postanowiłem poddać te zmienne dla których  $p$  w Chi kwadrat  $> 0.05$  regresji logistycznej ( funkcja `glm()` ). **Co Pan o tym myśli?**

```
modelowyframe <- c(3, 24) #dodaje do wektora numer kolumny zmiennej zależnej (nielegal_ogol), żeby poter
danelog <- subset(dane, select = c(modelowyframe))
```

```
model <-
  glm(nielegal_ogol ~ .,
      family = binomial(link = 'logit'),
      data = danelog)

summary(model)
```

```
##
## Call:
## glm(formula = nielegal_ogol ~ ., family = binomial(link = "logit"),
```

```

##      data = danelog)
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -0.4956  -0.3384  -0.3384  -0.3384   2.4036
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.0350     0.2217  -9.178  <2e-16 ***
## `M=0/K=1`    -0.7964     0.2850  -2.794   0.0052 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 404.66  on 791  degrees of freedom
## Residual deviance: 397.27  on 790  degrees of freedom
## AIC: 401.27
##
## Number of Fisher Scoring iterations: 5
modelowyframe <- c(istotnenummer, 24) #dodaje do wektora numer kolumny zmiennej zależnej (nielegal_ogol)
danelog <- subset(dane, select = c(modelowyframe))

model <-
  glm(nielegal_ogol ~ .,
      family = binomial(link = 'logit'),
      data = danelog)

summary(model)

##
## Call:
## glm(formula = nielegal_ogol ~ ., family = binomial(link = "logit"),
##      data = danelog)
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -2.5444  -0.3208  -0.1855  -0.1397   3.0890
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.35954     1.56791  -2.143  0.03214 *
## Wiek           -0.00628     0.06527  -0.096  0.92335
## `M=0/K=1`     -0.14933     0.34115  -0.438  0.66158
## Namawianie      0.08983     0.12435   0.722  0.47003
## Ciekawe        -0.18097     0.14957  -1.210  0.22629
## Kierunek_medyczny -0.74558     0.33389  -2.233  0.02555 *
## Kofeina         0.40252     0.52093   0.773  0.43971
## Nikotyna        1.74012     0.37673   4.619 3.86e-06 ***
## alkohol         1.45237     0.34362   4.227 2.37e-05 ***
## napoje_energetyzujące 0.02396     0.38372   0.062  0.95021
## dopalacze       3.16188     1.11219   2.843  0.00447 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 404.66  on 791  degrees of freedom
## Residual deviance: 286.46  on 781  degrees of freedom
## AIC: 308.46
##
## Number of Fisher Scoring iterations: 6
```

Zatem gdy brać pod uwagę wszystkie zmienne to jedynie alkohol, nikotyna i dopalacze są istotnymi statystycznie predyktorami (pozytywnie przewdującymi zmienną zależną). Natomiast gdyby wykluczyć je z modelu:

```
nikotyna_alkohol <- c(13, 14, 15, 16) #tworze wektor z numerami kolumn alkoholu, nikotyny i dopalaczy t
modelowyframe <- setdiff(istotnenum, nikotyna_alkohol) #odejmuje go od numerów zmiennych które wrzuc
modelowyframe <- c(modelowyframe, 24) #dodaje do wektora numer kolumny zmiennej zależnej (nielegal_ogol
danelog <- subset(dane, select = c(modelowyframe))

model <-
  glm(nielegal_ogol ~ .,
      family = binomial(link = 'logit'),
      data = danelog)

summary(model)
```

```
##
## Call:
## glm(formula = nielegal_ogol ~ ., family = binomial(link = "logit"),
##      data = danelog)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.1336  -0.4102  -0.3118  -0.2262   2.9137
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.05771    1.25599  -3.231  0.00123 **
## Wiek           0.07395    0.04839   1.528  0.12646
## `M=0/K=1`    -0.75444    0.29483  -2.559  0.01050 *
## Namawianie     0.26082    0.10860   2.402  0.01632 *
## Ciekawe       -0.21683    0.13809  -1.570  0.11635
## Kierunek_medyczny -0.85727    0.30158  -2.843  0.00447 **
## Kofeina        1.31368    0.44813   2.931  0.00337 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 404.66  on 791  degrees of freedom
## Residual deviance: 366.28  on 785  degrees of freedom
## AIC: 380.28
```



```
##  
## Number of Fisher Scoring iterations: 6
```

Okazuje się że kierunek medyczny jest najsilnie niezależnym predyktorem jest studiowanie na kierunku medyczny, a co najważniejsze przewiduje on negatywnie zmienną niezależną.

Odds Ratio i odpowiednie przedziały ufności wyglądają następująco:

```
exp(cbind(OR = coef(model), confint(model))) %>% round(3)
```

```
## Waiting for profiling to be done...  
  
##                OR 2.5 % 97.5 %  
## (Intercept)    0.017 0.001  0.219  
## Wiek           1.077 0.970  1.177  
## `M=0/K=1`     0.470 0.265  0.847  
## Namawianie     1.298 1.045  1.602  
## Ciekawe        0.805 0.616  1.060  
## Kierunek_medyczny 0.424 0.231  0.759  
## Kofeina        3.720 1.666  9.943
```

## Odważne wnioski:

W badanej grupie fakt bycia na studiach medycznych był samodzielnym (negatywnym) predyktorem używania substancji nielegalnych.

Płeć męska oraz większe ‘deklarowane bycie namawianym do studiowania’ były pozytywnymi predyktorami.