Analiza Danych Ankietowych - Raport 2.

Opracowanie zadań z listy 2.

Katarzyna Karbowska	268879
Maciej Ostapiuk	268810

19 maja 2024

1 Wstęp

1.1 Opis danych

Ten sam zestaw danych co w poprzednim raporcie został poddany analizie. Dane dotyczą ankiety przeprowadzonej w pewnej dużej agencji reklamowej, która miała na celu ocenę satysfakcji z pracy. Wiadomo że w ankiecie wzięło udział dwieście losowo wybranych osób. W tym zestawie znajduje się osiem następujących zmiennych, będących odpowiedziami na poszczególne pytania:

- **DZIAŁ** - jest odpowiedzią na pytanie ”W jakim dziale jesteś zatrudniony”, przyjmuje wartości HR (Dział obsługi kadrowo-płacowej), IT (Dział utrzymania sieci i systemów informatycznych), DK (Dział Kreatywny) lub DS (Dział Strategii)
- **STAŻ** - zmienna odpowiadająca na pytanie ”Jak długo pracujesz w firmie?”, przyjmująca wartości 1 (poniżej 1 roku), 2 (Między jednym rokiem a trzema latami), 3 (Powyżej trzech lat)
- **CZY_KIER** - zmienna będąca odpowiedzią na pytanie ”Czy pracujesz na stanowisku menedżerskim?”, przyjmująca wartości Tak (osoba jest na stanowisku menedżerskim) lub Nie (osoba obejmuje inne stanowisko niż menedżerskie)
- **PYT_1** - stanowi odpowiedź na pytanie ”Jak bardzo zgadzasz się ze stwierdzeniem, że firma pozwala na elastyczne godziny pracy tym samym umożliwiając zachowanie równowagi między pracą a życiem prywatnym?”, przyjmuje wartości: -2 (zdecydowanie się nie zgadzam), -1 (nie zgadzam się), 0 (nie mam zdania), 1 (zgadzam się), 2 (zdecydowanie się zgadzam)
- **PYT_2** - odpowiada zadane pytanie ”Jak bardzo zgadzasz się ze stwierdzeniem, że twoje wynagrodzenie adekwatnie odzwierciedla zakres wykonywanych przez ciebie obowiązków? i przyjmuje odpowiednie wartości -2 (zdecydowanie się nie zgadzam), -1 (nie zgadzam się), 1 (zgadzam się), 2 (zgadzam się)
- **WIEK** - oznacza wiek respondenta
- **PŁEĆ** - wskazuje na płeć ankietowanego
- **PYT_3** - jest odpowiedzią dotyczącą wynagrodzenia z pracy, po rewizji wynagrodzeń, w której część pracowników otrzymało podwyżki.

1.2 Wczytanie danych

Podobnie jak w poprzednim raporcie dokonaliśmy odpowiednich przekształceń na naszym zbiorze danych, zapewniając odpowiednie typy zmiennych oraz dodając niezbędne zmienne kategoryczne wymagane do dalszej analizy. Poniżej znajduje się kod, który był odpowiedzialny za te działania. Należy wspomnieć, że zmienna CZY_ZADOW łączy zmienną PYT_2 a zmienna CZY_ZADOW_2 zmienną PYT_3. Przyjmowaliśmy, że przyjmują one odpowiednie wartości 0 gdy osoba jest nie zadowolona oraz 1 gdy osoba jest zadowolona z wynagrodzenia odpowiednio w pierwszym badanym okresie oraz w drugim badanym okresie.

```r
data <- read.csv("ankieta.csv", sep = ';',
        col.names = c('DZIAL', 'STAZ', 'CZY_KIER', 'PYT_1', 'PYT_2', 'PYT_3',
        'PLEC', 'WIEK'))

data <- data %>%
  mutate_at(vars(DZIAL, STAZ, CZY_KIER, PYT_1, PYT_2, PYT_3, PLEC), as.factor)

przedzialy_wiekowe <- c(0, 35, 45, 55, Inf)
nazwy_kategori <- c("0-35", "36-45", "46-55", "55+")
data$WIEK_KAT <- cut(data$WIEK, przedzialy_wiekowe,
                    labels = nazwy_kategori,
                    include.lowest = TRUE)

data <- mutate(data,
    CZY_ZADOW = ifelse(as.numeric(as.character(PYT_2)) == -2, 0,
    ifelse(as.numeric(as.character(PYT_2)) == -1, 0,
        ifelse(as.numeric(as.character(PYT_2)) == 1, 1,
    ifelse(as.numeric(as.character(PYT_2)) == 2, 1, "_")))))

data$CZY_ZADOW <- as.factor(as.integer(data$CZY_ZADOW))

data <- mutate(data,
    CZY_ZADOW_2 = ifelse(as.numeric(as.character(PYT_3)) == -2, 0,
    ifelse(as.numeric(as.character(PYT_3)) == -1, 0,
        ifelse(as.numeric(as.character(PYT_3)) == 1, 1,
    ifelse(as.numeric(as.character(PYT_3)) == 2, 1, "_")))))

data$CZY_ZADOW_2 <- as.factor(as.integer(data$CZY_ZADOW_2))
```

2 Część I

2.1 Zadanie 1

W ankiecie przedstawionej na poprzedniej liście pracownicy zostali poproszeni o wyrażenie opinii na temat podejścia firmy do utrzymania równowagi między życiem zawodowym a prywatnym. Wsród próbki 200 pracowników (losowanie proste ze zwracaniem) uzyskano wyniki:

- **14 pracowników - bardzo niezadowolonych,**
- **17 pracowników - niezadowolonych,**
- **40 pracowników - nie ma zdania,**
- **100 pracowników - zadowolonych,**
- **29 pracowników - bardzo zadowolonych,**

Na podstawie danych wyznacz przedział ufności dla wektora prawodobieństw opisującego stopień zadowolenia z podejścia firmy. Przyjmij poziom ufności 0.95
Wykonaliśmy przedziały ufności dla rozkładu dwumianowego, biorąc każdą ze współrzędnych wejściowego wektora jako zmienną z rozkładu $\mathcal{B}(n, p_i), i = 1, ..., 5$. Przedziały zostały wyznaczone w wersjach: dokładnej i asymptotycznej. Dla porównania wykonaliśmy również, z pomocą

pakietu `multinomCI`, przedziały jednoczesne dla roważanego wektora. W tej metodologii stosujemy poprawkę Bonferroniego, gdzie dla i-tej współrzędnej wektora przyjmujemy poziom ufności $1 - \alpha_i = 1 - \frac{\alpha}{5}$.

```
responses <-c("very dissatisfied"=14, "dissatisfied"=17, "neither sattisfied nor
    dissatisfied"=40, "satisfied" = 100, "highly satisfied" = 29)
res_binom_ci <- binom.confint(responses, c(200, 200, 200, 200, 200), conf.level
    = 1-0.05/5, methods = c("exact", "asymptotic"))
res_multinom_ci <- multinomialCI(responses, alpha = 0.05/5)
```

Wyniki umieściliśmy w tabeli.

Odpowiedź	$\hat{p}_i$	**Dokładne**	**Asymptotyczne**	**Jednoczesne**
Bardzo niezadowolony	0.070	$[0.031, 0.129]$	$[0.023, 0.116]$	$[0.000, 0.158]$
Niezadowolony	0.085	$[0.042, 0.148]$	$[0.034, 0.135]$	$[0.000, 0.173]$
Nie ma zdania	0.200	$[0.132, 0.282]$	$[0.127, 0.272]$	$[0.115, 0.288]$
Zadowolony	0.500	$[0.407, 0.592]$	$[0.408, 0.569]$	$[0.415, 0.588]$
Bardzo zadwowolony	0.145	$[0.087, 0.221]$	$[0.081, 0.209]$	$[0.060, 0.233]$

Tabela 1: Przedziały ufności dla wyrażonej opinii.

Tabela jasno wskazuje na dwie najbardziej prawdopodobne odpowiedzi przy losowaniu pracowników- w większości przypadków pracownik będzie zadowolony bądź nie będzie potrafił wskazać, czy jest zadowolony. Możemy też zauważyć, że uzyskane przedziały są dosyć szerokie. Ma to związek z tym, że metodologia obliczania jednoczesnych przedziałów asymptotycznie dąży do poziomu ufności $1 - \alpha$. Przy obliczaniu przedziałów trzeba stosować poprawkę Bonferroniego, co przy większej ilości współrzędnych wektora $\mathbf{p}$ może wymagać większej ilości obserwacji w celu uzyskania asymptotycznych własności.

2.2 Zadanie 2

Napisz funkcję, która wyznacza wartość poziomu krytycznego w następujących testach:

- **chi-kwadrat Pearsona**
- **chi-kwadrat największej wiarogodności**

służących do weryfikacji hipotezy $H_0 : \mathbf{p} = \mathbf{p_0}$ przy hipotezie alternatywnej $H_1 : \mathbf{p} \neq \mathbf{p_0}$, na podstawie obserwacji $\mathbf{x}$ wektora losowego $\mathbf{X}$ z rozkładu wielomianowego z parametrami n i $\mathbf{p}$.

```
get_p_value <-function(x,n,h_0){
  #' Calculates p_values of user's choice test: either chi-square or maximum-
    likelihood chi-square
  sample <-  rmultinom(1, sum(x), h_0)
  point_probs_estimator = sample/n

  chi_square_test_statistic <- sum((x - n*h_0)^2/(n*h_0))
  chi_square_maximum_likelihood_test_statistic <- sum((x - n*point_probs_
    estimator)^2/(n*point_probs_estimator))
  k <-  length(x)

```

```
  p_value_chi_sq <- 1-pchisq(chi_square_test_statistic, df = k-1)

  p_value_chi_sq_max_likelihood <- 1-pchisq(chi_square_maximum_likelihood_test_statistic, df = k-1)

  return(c("p_value of chi-square test" = p_value_chi_sq, "p_value of maximum-likelihood chi-square test" = p_value_chi_sq_max_likelihood ))
}

sample_multinom <-  rmultinom(1, 100, c(0.1, 0.2, 0.4, 0.2, 0.1))

null_hypothesis <-get_p_value(x = sample_multinom, n = 100, h_0 = c(0.1, 0.2, 0.4, 0.2, 0.1))

alternative_hypothesis <-  get_p_value(x = sample_multinom, n = 100, h_0 = c(0.1, 0.2, 0.2, 0.4, 0.1))
```

Funkcja bazuje na obliczaniu statystyki testowej w postaci:

$$\chi^2(X) = \sum_{i=1}^{k} \frac{(X_i - np_{0i})^2}{np_{0i}},$$

gdzie n to parametr rozkładu $\mathcal{M}(n, \mathbf{p})$, a p_{0i} to elementy $\mathbf{p_0}$. W wersji największej wiarygodności, zamieniamy parametry p_{0i} punktowymi estymatorami NW, podanymi na wykładzie. W tym celu, generujemy próbkę z rozkładu podanego przez użytkownika jako argumenty wejściowe funkcji n, h_0. Testowanie wykonaliśmy na dwa sposoby- najpierw rozważyliśmy próbkę zgodną z rozkładem podanym w hipotezie zerowej (zmienna `nullhypothesis`)- test chi-kwadrat Pearsona zwrócił p-wartość w przybliżeniu równą 0.30, natomiast wersja chi-kwadrat największej wiarygodności zwróciła p-wartość równą 0.35. Zgodnie z oczekiwaniami, wyniki nie dostarczyły nam powodów do odrzucenia H_0. Drugi przypadek testowania polegał na wylosowaniu próbki z innego rozkładu niż ten założony w H_0. Różnica polegała na innym wektorze prawdopodobieństw. Zgodnie z założeniami, p-wartości testów chi-kwadrat Pearsona i chi-kwadrat najwiekszej wiarygodności wyniosły odpowiednio: $3.02 \cdot 10^{-6}$ oraz $2.37 \cdot 10^{-7}$.

2.3 Zadanie 3

Na podstawie danych z ankiety z poprzedniej listy zweryfikuj hipotezę, że w grupie pracowników zatrudnionych w Dziale Kreatywnym rozkład odpowiedzi na pytanie dotyczące podejścia firmy do utrzymania równowagi między życiem zawodowym a prywatnym jest równomierny, tzn. jest jednakowe prawdopodobieństwo, że pracownik zatrudniony w Dziale Kreatywnym udzielił odpowiedzi "zdecydowanie się nie zgadzam", "nie zgadzam się", "nie mam zdania", "zgadzam się", "zdecydowanie się zgadzam"na pytanie PYT_1. Przyjmij poziom istotności 0.05. Skorzystaj z funkcji napisanej w zadaniu 2.

Jeżeli chcemy testować rozkład równomierny, to tak naprawdę wektor $\mathbf{p_0}$ z hipotezy zerowej musi mieć wartość na każdej współrzędnej równą $\frac{1}{k}$. Zatem, testowaliśmy hipotezę $H_0 : \mathbf{p} = \mathbf{p_0} = (0.2, 0.2, 0.2, 0.2, 0.2)$ przeciwko alternatywie $H_1 : \mathbf{p} \neq \mathbf{p_0}$. Do przeprowadzenia testów użyliśmy funkcji napisanej w zadaniu 2.

```
counts <- data %>%
  filter(DZIAL == "DK") %>%
  group_by(PYT_1) %>%
```

```
  summarise(count = n())

creative_emps_responses <-  counts$count
print(creative_emps_responses)


get_p_value(creative_emps_responses, sum(creative_emps_responses), c(0.2, 0.2,
    0.2, 0.2, 0.2))
```

P-wartości w obydwu przypadkach są rzędu 10^{-13}, co odrzuca hipotezę zerową o równomiernym rozkładzie. Co więcej, aż 51 z 98 badanych pracowników wskazało odpowiedź "zgadzam się".

3 Część II

3.1 Zadanie 5

Korzystając z testu Fishera, na poziomie istotności 0.05 **, zweryfikuj hipotezę, że zmienna PŁEĆ i zmienna CZY_KIER są niezależne. Czy na poziomie istotności** 0.05 **możemy wnioskować, że prawdopodobieństwo tego, że na stanowisku kierowniczym pracuje kobieta jest równe prawdopodobienstwu tego, że na stanowisku kierowniczym pracuje mężczyzna?**

Dla tabeli dwudzielczej o wymiarze 2 x 2 test Fishera bada hipoteze zerową o niezależności zmiennych oraz dodatkowo sprawdza nam czy proporcje pomiędzy grupami są takie same. W zadaniu mamy sprawdzić hipotezę zerową czy zmienne PŁEĆ i CZY_KIER są niezależne, oraz czy prawdopodobieństwo, że na stanowisku pracowniczym pracuje kobieta jest równe że na stanowisku kierowniczym pracuje mężczyzna. Zastosowanie testu Fishera zamieszczone jest poniżej w kodzie.

```
ftable_PLEC_CZY_KIER <- table('Plec' = data$PLEC,
    'Stanowisko kierownicze' = data$CZY_KIER)

fisher_test <- fisher.test(ftable_PLEC_CZY_KIER, conf.int = TRUE)

cat('Przedzialy ufnosci:',fisher_test$conf.int)
cat('P-wartosc:',fisher_test$p.value)
```

Wykonany test na poziomie istotności 0.05 wskazuje p-wartość 0.6659029, zatem nie ma podstaw do odrzucenia hipotezy zerowej. (Co więcej przedziały ufności to [0.5299411, 3.802304].) Mianowicie nie możemy powiedzieć, że bycie na stanowisku kierownicznym zależy od płci oraz że proporcje pomiędzy płcią na stanowisku kierowniczym są nie równe.

3.2 Zadanie 6

Korzystajac z testu Freemana-Haltona na poziomie istotności 0.05 **zweryfikuj następujące hipotezy:**

1. **zajmowanie stanowiska kierowniczego nie zalezy od wieku (CZY_KIER oraz WIEK_KAT),**
2. **zajmowanie stanowiska kierowniczego nie zależy od stażu pracy (CZY_KIER oraz STAŻ),**

3. **zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od zajmowanego stanowiska (PYT_2 oraz CZY_KIER),**

4. **zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od stażu (PYT_2 oraz STAŻ),**

5. **zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od płci (PYT_2 oraz PŁEĆ)**

6. **zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od wieku (PYT_2 oraz WIEK_KAT).**

Powyższe hipotezy zostały zweryfikowane testem Freemana-Haltona, który również kryje się pod funkcją *fisher.test* w języku R.

1. **zajmowanie stanowiska kierowniczego nie zalezy od wieku (CZY_KIER oraz WIEK_KAT),**

```
ftable_CZY_KIER_WIEK_KAT <- table('Stanowisko kierownicze' = data$CZY_KIER,
    'Kategoria wiekowa' = data$WIEK_KAT)

fisher_test_1 <- fisher.test(ftable_CZY_KIER_WIEK_KAT, conf.int = TRUE)
cat('P-wartosc:',fisher_test_1$p.value)
```

   Na początku rozważaliśmy tezę, że zajmowanie stanowiska kierowniczego nie jest związane z wiekiem. Warto zauważyć, że test przy poziomie istotności 0.05 nie dostarczył wystarczających dowodów na odrzucenie hipotezy zerowej. P-wartość wyniosła 0.7823, co oznacza, że przekracza ona poziom istotności, sugerując brak istotności statystycznej.

2. **zajmowanie stanowiska kierowniczego nie zależy od stażu pracy (CZY_KIER oraz STAŻ),**

```
ftable_CZY_KIER_STAZ <- table('Stanowisko kierownicze' = data$CZY_KIER,
    'Dlugosc stazu pracy' = data$STAZ)

fisher_test_2 <- fisher.test(ftable_CZY_KIER_STAZ, conf.int = TRUE)
cat('P-wartosc:',fisher_test_2$p.value)
```

   Kolejna hipoza zerowa zakłada że zajmowanie stanowiska kierowniczego nie zależy od stażu pracy. Zweryfikowaliśmy tą hipotezę, podobnie tak jak wcześniej testem Freedmana-Haltona dla tablic o większych wymiarach niż 2x2. P-wartość w wykonanym teście jest równa $6.538 \cdot 10^{-5}$ co zaprzecza hipotezie zerowej, co jest dość intuicyjne, że zajmowanie stanowiska kierowniczego zależy o długości stażu pracy.

3. **zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od zajmowanego stanowiska (PYT_2 oraz CZY_KIER),**

```
ftable_PYT2_CZY_KIER <- table('Poziom zadowolenia w 1. badanym okresie' =
   data$PYT_2,
    'Stanowisko kierownicze' = data$CZY_KIER)

fisher_test_3 <- fisher.test(ftable_PYT2_CZY_KIER, conf.int = TRUE)
cat('P-wartosc:',fisher_test_3$p.value)
```

Badaliśmy hipotezę zerową sugerującą, że zadowolenie z wynagrodzenia w pierwszym okresie badawczym nie jest związane z zajmowanym stanowiskiem. Wynik testu Freedmana-Haltona wykazał p-wartość wynoszącą 0.0443, co jest niższe od ustalonego poziomu istotności 0.05. W związku z tym istnieją podstawy do odrzucenia hipotezy zerowej. Warto zauważyć, że zwykle osoby na stanowiskach kierowniczych otrzymują wyższe wynagrodzenie, co może przyczynić się do większego zadowolenia z poziomu wynagrodzenia.

4. **zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od stażu (PYT_2 oraz STAŻ),**

```
ftable_PYT2_STAZ <- table('Poziom zadowolenia w 1. badanym okresie' = data$
    PYT_2,
     'Dlugosc stazu pracy' = data$STAZ)

fisher_test_4 <-fisher.test(ftable_PYT2_STAZ, conf.int = TRUE)
cat('P-wartosc:',fisher_test_4$p.value)
```

Testowaliśmy hipotezę zerową która mówi, że zadowolenie w pierwszym badanym okresie nie zalezy od stazu pracy Wykonany test Freedmana-Haltona wykazał p-wartość 0.01069. co jest mniejsze od hipotezy zerowej, więc istenieją podstawy na odrzucenie tezy że zadowolenie w pierwszym badanym okresie nie zalezy od stazu pracy. Osoby z dłuższym stażem mogą mieć większe doświadczenie i umiejętności, co często przekłada się na wyższe zarobki i potencjalnie większe zadowolenie z wynagrodzenia. Ponadto, wraz z upływem czasu pracownik może zdobywać awanse lub dodatkowe kwalifikacje, co również może wpłynąć na jego poziom zadowolenia z wynagrodzenia. Jednakże zależność ta może być różna w zależności od wielu czynników, takich jak branża, lokalizacja, polityka płacowa firmy, czy nawet osobiste preferencje i wartości pracownika.

5. **zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od płci (PYT_2 oraz PŁEĆ)**

```
ftable_PYT2_PLEC <- table('Poziom zadowolenia w 1. badanym okresie' = data$
    PYT_2,
     'Plec' = data$PLEC)

fisher_test_5 <- fisher.test(ftable_PYT2_PLEC, conf.int = TRUE)
cat('P-wartosc:',fisher_test_5$p.value)
```

Wynik testu Freedmana-Haltona wskazuje, że p-wartość wynosi 0.4758, co jest znacznie wyższe od przyjętego poziomu istotności $\alpha = 0.05$. Oznacza to, że nie ma wystarczających dowodów na odrzucenie hipotezy zerowej, która zakłada, że zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od płci. Innymi słowy, na podstawie tego testu nie możemy stwierdzić istotnego związku między płcią a zadowoleniem z wynagrodzenia w pierwszym badanym okresie.

6. **zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od wieku (PYT_2 oraz WIEK_KAT).**

```
ftable_PYT2_WIEK_KAT <- table('Poziom zadowolenia w 1. badanym okresie' =
    data$PYT_2, 'Kategoria wiekowa' = data$WIEK_KAT)

fisher_test_6 <- fisher.test(ftable_PYT2_WIEK_KAT, conf.int = TRUE,
    workspace=500000)
cat('P-wartosc:',fisher_test_6$p.value)
```

Warto zauważyć, że wynik testu Freedmana-Haltona wykazał p-wartość równą 0.3194, co jest wyższe od ustalonego poziomu istotności $\alpha = 0.05$. Oznacza to, że nie ma wystarczających dowodów na odrzucenie hipotezy zerowej, która zakłada, że zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od wieku. To przekonanie wydaje się uzasadnione, ponieważ firmy często klasyfikują wynagrodzenia na podstawie kryteriów takich jak wykształcenie, umiejętności czy stanowisko zajmowane przez pracownika.

4 Część III

4.1 Zadanie 8

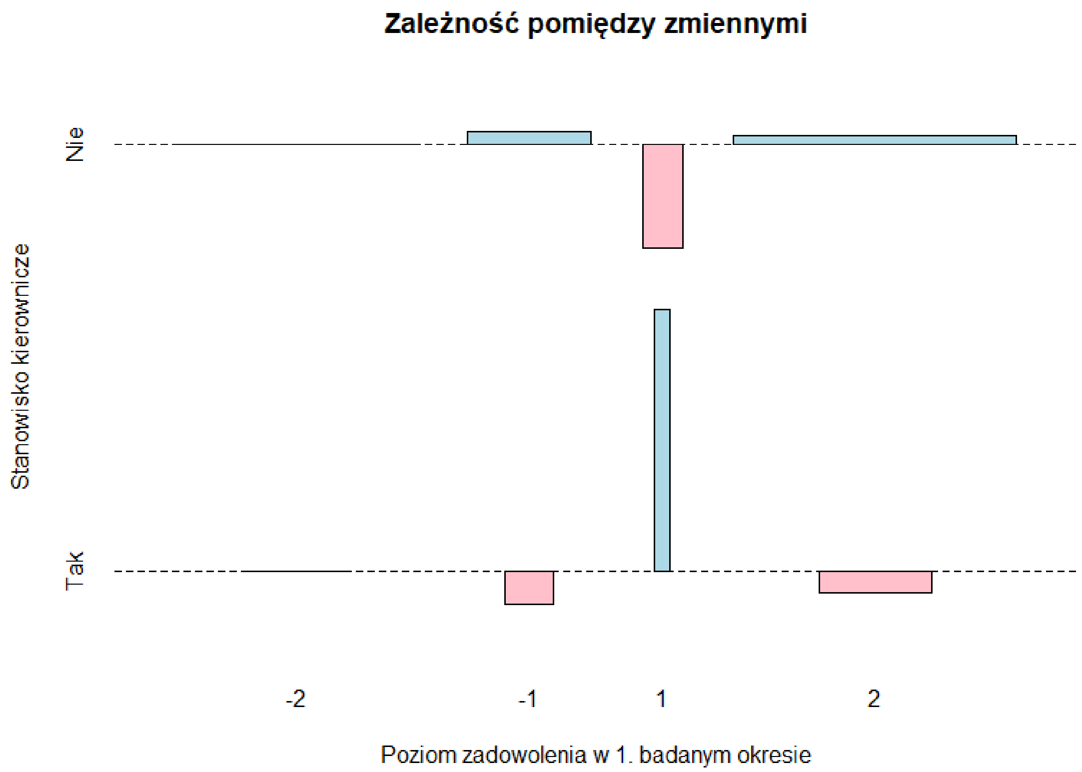
Korzystając z *chisq.test* zweryfikuj hipotezę, że zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od zajmowanego stanowiska. Przyjmij poziom istotności 0.01. Stwórz wykres przy pomocy funkcji *assocplot* i dokonaj jego interpretacji. Wynik testu porównaj z wynikiem uzyskanym w zadaniu 6.

```
chisq_test_PYT2_CZY_KIER <- chisq.test(ftable_PYT2_CZY_KIER)

cat('P-wartosc:',chisq_test_PYT2_CZY_KIER$p.value)
```

W tym zadaniu ponownie przetestowaliśmy hipotezę zerową, sugerującą brak zależności między zadowoleniem z wynagrodzenia a zajmowanym stanowiskiem w pierwszym badanym okresie. Tym razem zdecydowaliśmy się wykorzystać test chi-kwadrat do weryfikacji tej hipotezy. Wynik testu w postaci p-wartości wynoszącej 0.004397 wykazał, że przy założonym poziomie istotności 0.01 istnieją wystarczające podstawy do odrzucenia hipotezy zerowej. Porównując test chi-kwadrat z testem Freedmana-Haltona (zadanie 6. test 3.) (p-wartość 0.0443), oba wykazują odrzucenie hipotezy zerowej na poziomie istotności 0.05. Jednakże tylko test chi-kwadrat utrzymuje to odrzucenie na poziomie istotności 0.01.

Wykonaliśmy również wykres asocjacji aby zidentyfikować wzorce i relacje pomiędzy tymi zmiennymi.

```
assocplot(ftable_PYT2_CZY_KIER,
    col = c("lightblue", "pink"),
    main = 'Zaleznosc pomiedzy zmiennymi',
    xlab = "Poziom zadowolenia w 1. badanym okresie",
    ylab = "Stanowisko kierownicze")
```



Rysunek 1: Wykres asocjacji z zadania 8.

Na wykresie asocjacji zauważamy, że prostokąty nie są równomiernie rozłożone. Ich różne wielkości sugerują, że istnieje związek między zmiennymi. Wartości prostokątów są znacząco różne, co wskazuje na nielosowy charakter występowania danych kategorii zmiennych. Jednakże, na podstawie samej różnorodności wielkości prostokątów nie możemy jednoznacznie stwierdzić, czy zmienne są niezależne. Konieczne jest przeprowadzenie dodatkowych analiz, takich jak test statystyczny, co wykonaliśmy wcześniej i potwierdziło to że zmienne są najprawdopodobniej zależne.

4.2 Zadanie 9

Zapoznaj się z funkcją *rmultinom* z pakietu stats, a następnie korzystając z niej przeprowadź symulacje w celu oszacowania mocy testu Fishera oraz mocy testu chi-kwadrat Pearsona, generując dane z tabeli 2 × 2, w której $p_{11} = \frac{1}{40}$, $p_{12} = \frac{3}{40}$, $p_{21} = \frac{19}{40}$, $p_{22} = \frac{17}{40}$. Symulacje wykonaj dla n = 50, n = 100 oraz n = 1000.

```
set.seed(12345)

simulate_power <- function(MC=500, alpha, n, p){
  power_fisher <- rep(0, MC)
  power_chisq <- rep(0, MC)
  power_chisq_uncorrect <- rep(0, MC)

  for (i in 1:MC) {
    multi <- rmultinom(n=1, size=n, prob=p)

    fisher_test <- fisher.test(matrix(multi, nrow=2))
```

```
    power_fisher[i] <- as.numeric(fisher_test$p.value < alpha)

    chisq_test <- chisq.test(matrix(multi+0.00000001, nrow=2), correct = TRUE)
    power_chisq[i] <- as.numeric(chisq_test$p.value < alpha)

    chisq_test_uncorrect <- chisq.test(matrix(multi+0.00000001, nrow=2),
       simulate.p.value = TRUE)
    power_chisq_uncorrect[i] <- as.numeric(chisq_test_uncorrect$p.value < alpha)
  }

  return(c(mean(power_fisher), mean(power_chisq), mean(power_chisq_uncorrect)))
}

p <- c(1/40, 3/40, 19/40, 17/40)
# dla alpha 0.05
power_50 <- simulate_power(MC=500, alpha=0.05, n=50, p)
power_100 <- simulate_power(MC=500, alpha=0.05, n=100, p)
power_1000 <- simulate_power(MC=500, alpha=0.05, n=1000, p)

df_05 <- data.frame('Typ testu' = c('test Fishera', 'test Chi2 z poprawka', '
   test Chi2 bez poprawki'), 'n=50' = power_50, 'n=100' = power_100,
    'n=1000'= power_1000)
view(df_05)

# dla alpha 0.01
power_50_01 <- simulate_power(MC=500, alpha=0.01, n=50, p)
power_100_01 <- simulate_power(MC=500, alpha=0.01, n=100, p)
power_1000_01 <- simulate_power(MC=500, alpha=0.01, n=1000, p)

df_01 <- data.frame('Typ testu' = c('test Fishera', 'test Chi2 z poprawka', '
   test Chi2 bez poprawki'), 'n=50' = power_50_01, 'n=100' = power_100_01,
         'n=1000'= power_1000_01)
view(df_01)
```

Powyżej zamieszczono kod służący do symulacyjnego wyznaczenia mocy testów. Wykorzystaliśmy funkcję *rmultinom* do generowania danych o zadanych prawdopodobieństwach. Przeanalizowaliśmy moc testów dla $n \in \{50, 100, 1000\}$. Testy obejmowały test Fishera oraz test chi-kwadrat, zarówno z poprawką Yatesa, jak i bez niej. Aby wykonać test bez poprawki skorzystaliśmy z zależności że funkcja *chisq.test* wykonuje test bez poprawki jeśli opcję *simulate.p.value* ustawimy na *TRUE* (Oczywiście można byłoby ustawić opcję *correct* na *FALSE*) [1]. Dodatkowow żeby uniknąć problemów z obliczeniami przy zerach, dodaliśmy do każdej wartości wektora *multi* niewielką wartość. W poniższych tabelach znajdziemy wyniki wykonanania symulacji.

Rodzaj testu	$n = 50$	$n = 100$	$n = 1000$
test Fishera	0.112	0.298	1
test Chi2 z poprawka	0.086	0.242	1
test Chi2 bez poprawki	0.232	0.388	1

Tabela 2: Moce testów dla poziomu istotności $\alpha = 0.05$.

Rodzaj testu	$n = 50$	$n = 100$	$n = 1000$
test Fishera	0.030	0.114	0.992
test Chi2 z poprawka	0.014	0.066	0.990
test Chi2 bez poprawki	0.160	0.172	0.990

Tabela 3: Moce testów dla poziomu istotności $\alpha = 0.01$.

Zauważmy że, wraz ze wzrostem liczby obserwacji rośnie wartość mocy dla wszystkich wykonanych testów na każdym analizowanym poziomie istotności. Porównując poszczególne testy to dla testu Chi-kwadrat bez poprawki występują zdecydowanie większe wartości, zatem można przypuszczać że ten test jest najmocniejszy wśród rozważnych testów. Porównując natomiast oba poziomy istotności to widzimy że dla $\alpha = 0.05$ są to większe wartości. Podsumowując, poziom istotnotności oraz rozmiar próbki wpływają na moc testu.

4.3 Zadanie 10

Napisz funkcję, która dla danych z tablicy dwudzielczej oblicza wartość poziomu krytycznego w teście niezależnosci opartym na ilorazie wiarogodnosci. Korzystając z napisanej funkcji, wykonaj test dla danych z zadania 8.

Test oparty na ilorazie wiarogodności służy do weryfikacji hipotezy zerowej o niezależności zmiennych. Dla danych zadania 8 rozważaliśmy hipotezę zerową, która zakładała że, zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od zajmowanego stanowiska. Porównywaliśmy dwa poziomy istotności, dlatego wykonamy to samo w tym zadaniu. Poniżej funkcja wyznaczająca wartość krytyczną.

```
test_IW <- function(table){
  # liczba obserwacji
  n <- sum(table)
  C <- ncol(table)
  R <- nrow(table)

  # rozklady brzegowe
  R_totals <- rowSums(table)
  C_totals <- colSums(table)

  outer_table <- outer(R_totals, C_totals)

  expected <- (outer_table/(table*n))^table

  # statystyki
  lambda <- prod(expected)
  G2 <- -2*log(lambda)

  df <- (R-1)*(C-1)
  p_value <- 1-pchisq(G2, df)

  return (p_value)
}

cat('P-wartosc: ',test_IW(ftable_PYT2_CZY_KIER))
```

Test ilorazu wiarogodności dla danych analizowanych z zadania 8 wykazał p-wartość wynoszącą 0.03969. Jeśli weźmiemy pod uwagę poziom istotności 0.01 to nie ma podstaw do odrzucenia hipotezy zerowej, natomiast dla poziomu istotności 0.05 należy ją odrzucić.

5 Część IV

5.1 Zadanie 11

Przeprowadzone wsród brytyjskich mężczyzn badanie trwające 20 lat wykazało, że odsetek zmarłych (na rok) z powodu raka płuc wynosił 0.00140 **wsród osób palących papierosy i** 0.00010 **wsród osób niepalących. Odsetek zmarłych z powodu choroby niedokrwiennej serca wynosił** 0.00669 **dla palaczy i** 0.00413 **dla osób niepalących. Opisz związek pomiędzy paleniem papierosów a śmiercią z powodu raka płuc oraz związek pomiędzy paleniem papierosów a śmiercią z powodu choroby serca. Skorzystaj z różnicy proporcji, ryzyka względnego i ilorazu szans. Zinterpretuj wartości. Związek której pary zmiennych jest silniejszy?**
W zadaniu badamy związki między paleniem papierosów a śmiercią z powodu raka płuc oraz z powodu choroby serca. Poniżej widoczny jest kod, który pomógł w rozwiązaniu tego zadania.

```
smokers_tab <- matrix(c(0.00669,0.00140, 0.00413, 0.00010), nrow = 2, ncol = 2)

## ischemic heart disease and smoking
proportion_differences_disease_smoking = smokers_tab[1,1] - smokers_tab[1,2]
RR_disease_smoking = smokers_tab[1,1]/smokers_tab[1,2]
OR_disease_smoking <- RR_disease_smoking * (1 - smokers_tab[1,2])/ (1 - smokers_
    tab[1,1])


## lung cancer and smoking
proportion_differences_cancer_smoking<- smokers_tab[2,1] - smokers_tab[2,2]
RR_cancer_smoking<- smokers_tab[2,1]/smokers_tab[2,2]
OR_cancer_smoking <- RR_cancer_smoking * (1 - smokers_tab[2,2])/ (1 - smokers_
    tab[2,1])

results <- data.frame(
  Choroba = c("Choroba niedokrwienna serca", "Rak pluc"),
  `żRónica proporcji` = c(proportion_differences_disease_smoking,
                         proportion_differences_cancer_smoking),
  RR = c(RR_disease_smoking, RR_cancer_smoking),
  OR = c(OR_disease_smoking, OR_cancer_smoking)
)

xtable(results, digits = c(4, 4, 4, 4, 4))
```

Wyniki umieszczono w tabeli poniżej.

Choroba	Różnica.proporcji	RR	OR
Choroba niedokrwienna serca	0.0026	1.6199	1.6240
Rak płuc	0.0013	14.0000	14.0182

Tabela 4: Parametry opisujące związki pomiędzy paleniem papierosów a poszczególnymi chorobami

Różnica proporcji przedstawia różnice ryzyka zgonu między palaczami a niepalącymi. Dla raka płuc różnica proporcji wyniosła 0.0013, natomiast dla choroby niedokrwiennej serca ta różnica wyniosła 0.0026. Ryzyko względne (RR) pokazuje nam że palacze mają 14-krotnie wyższe ryzyko zgonu z powodu raka płuc oraz 1.6199-krotnie wyższe ryzyko zgonu z powodu choroby niedokrwiennej serca w porównaniu do osób niepalących. Iloraz szans (OR) wskazuje, że szansa na zgon z powodu raka płuc wśród palaczy jest 14.0182 a z powodu choroby niedokrwiennej serca jest 1.6240 razy większa niż dla niepalących. Analizując razem wszystkie wyniki, dochodzimy do wniosku że związek pomiędzy paleniem papierosów a śmiercią z powodu raka płuc jest większy niż związek pomiędzy paleniem śmiercią z powodu choroby niedokrwiennej serca. Palenie ma więc istotny wpływ na ryzyko śmierci z powodu raka płuc.

5.2 Zadanie 12

Tabela 1 przedstawia wyniki dotyczące śmiertelności kierowców i pasażerów w wypadkach samochodowych na Florydzie w 2008 roku, w zależności od tego, czy osoba miała zapięty pas bezpieczeństwa czy nie.

	Śmiertelny	**Nieśmiertelny**
Bez pasów	1085	$55,623$
Z pasami	703	$441,239$

Tabela 5: Wyniki dotyczące wypadków samochodowych na Florydzie w 2008 roku.

Do zadania został wykorzystany załączony kod:

```
accidents <- matrix(c(1085, 703, 55623, 441239), nrow = 2, ncol = 2)

sum_without_seatbelts <- 1085 + 55623
sum_seatbelts <- 703 + 441239
sum_mortal <-  1085 + 703
sum_immortal <- 55623 + 441239

prob_death_seatbelts = 703/sum_seatbelts
cat("Warunkowe prawdopodobienstwo smierci z wzgledu na pas:",
    prob_death_seatbelts)

prob_death_without_seatbelts = 1085/sum_without_seatbelts
cat("Warunkowe prawdopodobienstwo smierci z wzgledu na brak uzycia pasu:",
    prob_death_without_seatbelts)

prob_seatbelts_death <- 703/sum_mortal
cat("Warunkowe prawdopodobiezstwo uzycia pasu z wzgledu na wypadek smiertelny:",
    prob_seatbelts_death)

prob_seatbelts_live <- 441236/sum_immortal
cat("Warunkowe prawdopodobienstwo uzycia pasu z wzgledu na przezycie:",
    prob_seatbelts_live)
```

Przy pomocy tego kodu odpowiedzieliśmy na polecenia:

- **Oszacuj warunkowe prawdopodobieństwo śmierci w wypadku ze względu na drugą zmienną, tj. dla kierowców i pasażerów, który użyli pasa bezpieczeństwa oraz dla kierowców i pasażerów, który nie użyli pasa bezpieczeństwa.**

  Wartość tego prawdopodobieństwa wynosi 0.001590706 jeżeli chodzi o przypadki śmiertelne w sytuacji gdy pasażerowie i kierowca zapinają pasy bezpieczeństwa a 0.0191331 gdy rozważamy kierowców i pasażerów nie korzystających z pasów bezpieczeństwa. Zauważmy że dla przypadku w którym ludzie nie używają pasów prawdopodobieństwo jest większe, co sugeruje że ryzyko śmierci w wypadku jest wtedy wyższe.

- **Oszacuj warunkowe prawdopodobieństwo użycia pasa bezpieczeństwa ze względu na drugą zmienną, tj. dla kierowców i pasażerów ze śmiertelnymi obrażeniami oraz dla kierowców i pasażerów, którzy przeżyli wypadek.**

  Prawdopodobieństwo warunkowe użycia pasów bezpieczeństwa ze względu na wypadek śmiertelny wynosi 0.3931767 podczas gdy ze względu na pasażerow i kierowców którzy przeżyli wypadek wynosi 0.8880454. Możemy zatem strwierdzić że z tych którzy przeżyli wypadek miało pasy około 90%, więc pasy dają większe szanse na uniknięcie obrażeń śmiertelnych podczas wypadku.

- **Jaki jest najbardziej naturalny wybór dla zmiennej objaśnianej w tym badaniu? Dla takiego wyboru wyznacz i zinterpretuj różnicę proporcji, ryzyko względne oraz iloraz szans. Dlaczego wartości ryzyka względnego i ilorazu szans przyjmują zbliżone wartości?**

Najbardziej naturalny wybór jest wybrać zmienną objaśnianą jako śmiertelność wypadku - wtedy zmienna objaśniająca, czyli fakt zapięcia pasów bądź nie, pozwala ustalić wiele rzeczy, takich jak, przykładowo przyczyny zgonu.

```
proportion_differences_mortal<- prob_death_without_seatbelts - prob_death_
    seatbelts
RR_mortal <-  prob_death_without_seatbelts/prob_death_seatbelts
OR_mortal <-  RR_mortal * (1 - prob_death_seatbelts)/ (1 - prob_death_
    without_seatbelts)

cat("róznica proporcji: ", proportion_differences_mortal)
cat("RR mortal: ", RR_mortal)
cat("OR mortal: ", OR_mortal)

wyniki <- data.frame(
  "żRónica proporcji" = proportion_differences_mortal,
  "Ryzyko wzgledne (RR)" = RR_mortal,
  "Iloraz szans (OR)" = OR_mortal)
xtable(wyniki, digits = c(4, 4, 4, 4))
```

Wyniki umieszczone są w tabeli poniżej.

Różnica proporcji	Ryzyko względne (RR)	Iloraz szans (OR)
0.0175	12.0281	12.2432

Różnica proporcji wyniosła 0.0175, co oznacza że prawdopodobieństwo śmierci osób, które nie używają pasów bezpieczeństwa o tyle wyższe od prawdopodobieństwa śmierci osób używających pas. Ryzyko względne wraz z ilorazem szans przyjmują zbliżone wyniki bliskie 12. Te wyniki wyszły zbliżone, ponieważ odwołując się do relacji podanej na wykładzie:

$$OR = RR\frac{1-\pi_2}{1-\pi_1}$$

zauważmy, że zarówno π_1 - prawdopodobieństwo warunkowe śmierci ze względu na brak pasów, jak i π_2 - prawdopodobieństwo warunkowe śmierci ze względu na użycie pasu (z podpunktu pierwszego), były relatywnie blisko zera, co też oznacza to rzadkość występowania wypadków śmiertelnych.

W ogólności w sytuacjach o rzadkim wystepowaniu ryzyko względne i iloraz szans przyjmują zbliżone wartości, gdyż obie miary mają tendencję do zbiegania się do podobnych wartości.

5.3 Zadanie 13

Oblicz wartości odpowiednich miar współzmienności (współczynnik tau lub współczynnik gamma) dla zmiennych:

- **zadowolenie z wynagrodzenia w pierwszym badanym okresie i zajmowane stanowisko,**
- **zadowolenie z wynagrodzenia w pierwszym badanym okresie i staż pracy,**
- **zajmowane stanowisko i staż pracy.**

- Zadowolenie z wynagrodzenia w pierwszym badanym okresie i zajmowane stanowisko. Obie zmienne (`CZY_ZADW`, `CZY_KIER`) traktujemy jako nominalne, stąd obliczamy współczynnik τ.

```
czy_zadw_czy_kier_tau <- GoodmanKruskalTau(data$CZY_ZADOW,data$CZY_KIER)

```

  Wartość współczynnika wynosi około 0.0004, co oczywiście wynika z braku niezależności rozważanych zmiennych. Interpretacja jest następująca: stopień zmienności zadowolenia z wynagrodzenia w pierwszym badanym okresie przy zmienności zajmowanego stanowiska jest nieznaczny.

- Zadowolenie z wynagrodzenia w pierwszym badanym okresie i staż pracy.
  Są to zmienne porządkowe (kategoryczne), zatem wyliczyliśmy dla nich współczynnik γ.

```
pyt_2_staz_gamma <- GoodmanKruskalGamma(data$PYT_2, data$STAZ)

```

  Obliczona wartość wynosi w przybliżeniu 0.09084262, co po raz kolejny wskazuje na brak powiązania redukcji zmienności zadowolenia z wynagrodzenia w 1. badanym okresie ze zmiennością długości stażu.

- Zajmowane stanowisko i staż pracy.
  Mamy do czynienia z jedną zmienną nominalną (`CZY_KIER` - dychotomiczną) oraz z drugą zmienną porządkową. Dlatego przedstawimy interpretację tych dwóch zmiennych jako zmienne nominalne, przy czym dokonaliśmy transformacji zmiennej `STAZ` - stworzyliśmy zmienną `CZY_STAZ`, która przyjmuje wartość 1, gdy staż pracownika jest niedłuższy niż jeden rok, oraz 0 poza tym.

```
data <- mutate(data, CZY_STAZ = ifelse(as.numeric(STAZ) == 1, 0, 1))
czy_kier_staz_tau<- GoodmanKruskalTau(data$CZY_KIER, data$CZY_STAZ)
```

  Obliczona wartość wynosi w przybliżeniu 0.07, co świadczy o istnieniu słabej, bardzo znikomej relacji pomiędzy zmiennością zmiennej `CZY_KIER` oraz `STAZ`.

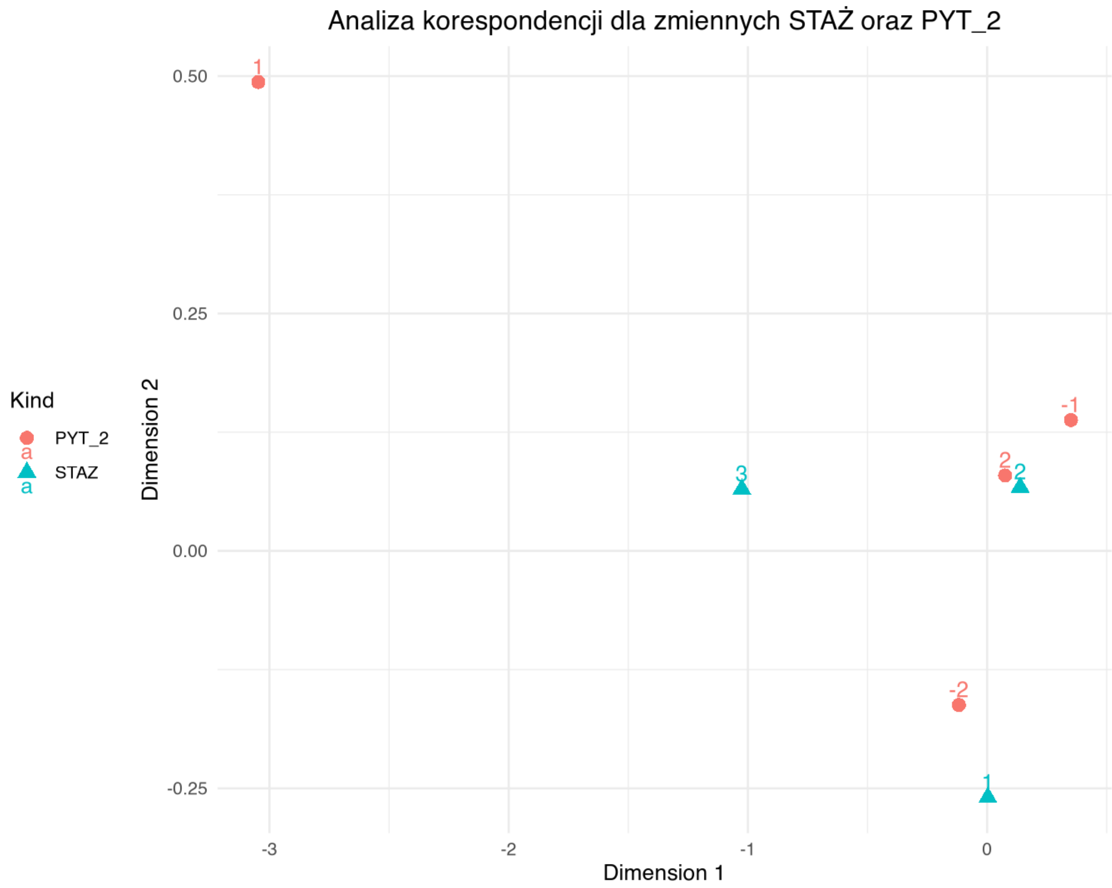
5.4 Zadanie 14

Na podstawie informacji przedstawionych na wykładzie napisz własną funkcję do przeprowadzania analizy korespondencji. Funkcja powinna przyjmować jako argument tablicę dwudzielczą i zwracać obliczone wartości odpowiednich wektorów i macierzy, współrzędnych punktów oraz odpowiedni wykres. Korzystając z napisanej funkcji wykonaj analizę korespondencji dla danych dotyczących zadowolenia z wynagrodzenia w pierwszym badanym okresie i stażu pracy.
W celu napisania funkcji posłużyliśmy się metodologią opisaną na wykładzie. Funkcja jako argument przyjmuje tablicę dwudzielczą, a zwraca macierze **F** i **G**, wykorzystywane do analizy principal-principal, wektory częstości brzegowych **r**, **c**, macierz profili wierszowych **R**, macierz profili kolumnowych **C**, macierz korespondencji **P** oraz wartość statystyki χ^2 wraz z inercją całkowitą λ. Wyniki zilustrowano na wykresie principal-principal oraz działanie funkcji porównano z funkcją `plot.ca` z modułu `ca`.

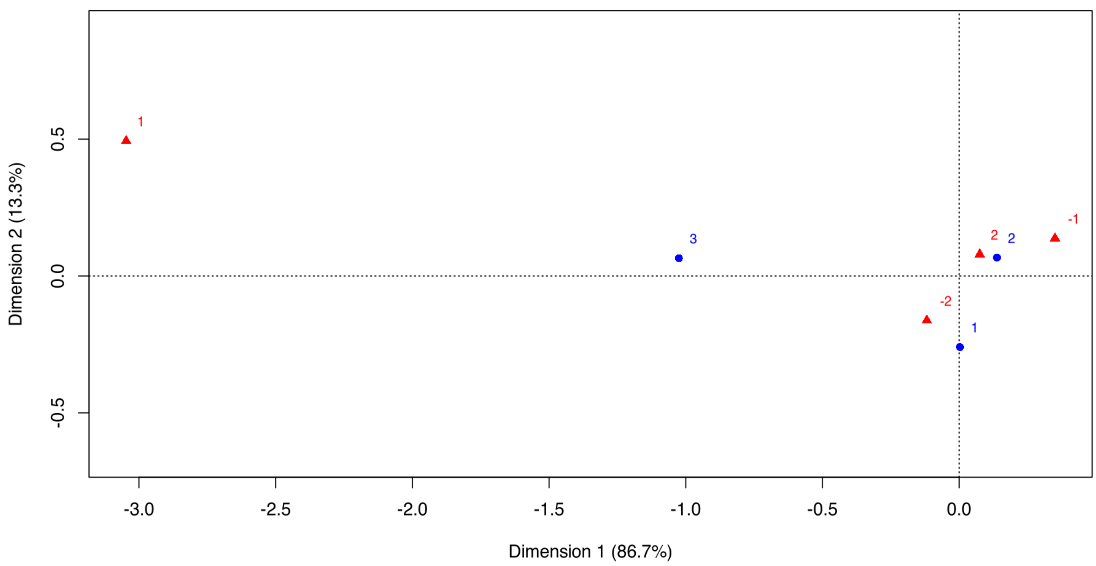
```
correspondence <- function(crosstab){
  N <- crosstab %>% as.matrix()
  P <-  N/sum(N)
```

```
  r <- P %>%  rowSums()
  c <- P %>%  colSums()
  D_r <-diag(r)
  D_c <-diag(c)
  R <-  solve(D_r)%*%P
  C <- P%*%solve(D_c)
  A <- solve(D_r)^(1/2)%*%(P - r%*%t(c))%*%solve(D_c)^(1/2)
  chi_sq <- sum(N) *sum(A^2)
  U <-  svd(A)$u
  Gamma_vec <-  svd(A)$d
  V <-  svd(A)$v
  nd <-  min(nrow(crosstab), ncol((crosstab))) - 1
  F_matrix <-  solve(D_r)^(1/2)%*%U[,1:nd]%*%diag(Gamma_vec[1:nd])
  G_matrix <-  solve(D_c)^(1/2)%*%V[,1:nd]%*%diag(Gamma_vec[1:nd])
  inertia <-  sum(diag(A%*%t(A)))
  return(list(F_matrix, G_matrix, r,c, R, C, P, inertia, chi_sq))
}

pyt_2_staz <- table(data[c("STAZ", "PYT_2")])

ca_pyt_2_staz <- correspondence(pyt_2_staz)
F_matrix <- ca_pyt_2_staz[[1]]
G_matrix <- ca_pyt_2_staz[[2]]

rows <-  data.frame("D1" = F_matrix[,1], "D2"=F_matrix[,2], "Value" = rownames(
    pyt_2_staz))
columns <-  data.frame("D1" = G_matrix[,1], "D2"=G_matrix[,2], "Value" =
    colnames(pyt_2_staz))
coordinates <- rbind( data.frame("D1" = rows$D1, "D2" =rows$D2, "Value" = rows$
    Value, "Kind" = "STAZ"), data.frame("D1" = columns$D1, "D2" = columns$D2, "
    Value" = columns$Value, "Kind" ="PYT_2"))

ggplot(coordinates, aes(x = D1, y = D2, color = Kind, shape = Kind, label =
    Value)) +
  geom_point(size = 3) +
  geom_text(vjust = -0.5) +
  theme_minimal() +
  labs(x = "Dimension 1", y = "Dimension 2") +
  ggtitle("Analiza korespondencji dla zmiennych ŻSTA oraz PYT_2")+
  theme(plot.title = element_text(hjust=0.5), legend.position = "left")

plot.ca(ca(pyt_2_staz, nd =2))
```

Podany kod wygenerował dwa wykresy:



Rysunek 2: Analiza korespondencji wykonana napisaną funkcją `correspondence`



Rysunek 3: Analiza korespondencji wykonana funkcją `plot.ca()` z modułu `ca`

Zauważmy, że wykresy są do siebie zbliżone, co wskazuje na poprawność napisanej przez nas funkcji. Dzięki tym wykresom (*principal-principal*), które określają podobieństwo profili jako ich odległość względem siebie na płaszczyźnie możemy zauważyć ciekawe obserwacje- Osoby pracujące dłużej niż 3 lata w rozważanej firmie zdecydowanie różnią się profilem od tych, którzy pracują co najwyżej dwa lata. Analogicznie, profil osób zadowolonych z wynagrodzenia w pierwszym badanym okresie znacznie różni się od profili osób skrajnie zadowolonych, bądź ogólnie niezadowolonych. Z tego powodu, wartośc inercji całkowitej układu złożonego z naszych profili jest różna od zera, wynosi w przybliżeniu 0.13, co wskazuje na istnienie rozproszenia pomiędzy profilami. Popierając to analizami wykonanymi na potrzeby poprzedniego i obecnego raportu, osoby w umiarkowanym stopniu zadowolone z wynagrodzenia w pierwszym badanym okresie są różnorodne pod względem zajmowanego stanowiska, działu, w którym pracują, płci itd. Inne grupy ze względu na odpowiedź na pytanie `PYT_2` mają zbliżone do siebie profile, co wskazywałoby na istnienie pewnych charakterystyk, które determinują ich skrajne zadowolenie, bądź niezadowolenie.

Literatura

[1] Data dostępu: 17.05.2024 r. (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/chisq.test)