Analiza danych ankietowych - sprawozdanie 2

"Lista zadań nr 2"

Przedmiot i prowadzący:
Analiza danych ankietowych,
poniedziałki 9.15 - 11.00 (grupa nr 2),
Inż. Hubert Woszczek

Aleksandra Hodera (268733) Aleksandra Polak (268786)

Spis treści

1 Wstęp

W poniższym sprawozdaniu zostały przedstawione wyniki listy zadań nr 2 przygotowanej w ramach laboratoriów z analizy danych ankietowych prowadzonych przez inż. Huberta Woszczek do wykładu dr inż. Aleksandry Grzesiek.

2 Użyte biblioteki

W tym punkcie zostały przedstawione wszystkie biblioteki, które użyłyśmy podczas tworzenia raportu:

```
library(ggplot2)
library(tidyr)
library(dplyr)
library(xtable)
library(binom)
library(reshape2)
library(stats)
library(graphics)
library(DescTools)
library(ca)
library(energy)
```

3 Wczytywanie danych

Zanim przeszłyśmy do wykonania zadań z listy wczytałyśmy dane z pliku o nazwie *"ankieta.csv"*.

```
data <- read.csv("ankieta.csv", fileEncoding = "Latin1", sep=";", na=c("")
    )

colnames(data) <- c('DZIAŁ','STAŻ','CZY_KIER', 'PYT_1', 'PYT_2', 'PYT_3',
     'PŁEĆ', 'WIEK')

data <- mutate(data, WIEK_KAT = cut(WIEK, breaks = c(0, 35, 45, 55, max(
    WIEK)), labels = c("0-35", "36-45", "46-55", "56+")))

data$CZY_ZADOW <- ifelse(data$PYT_2 %in% c("-2", "-1"), "NIE", "TAK")
```

4 Część I

4.1 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. Wśród 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 prawdopodobieństw opisującego stopień zadowolenia z podejścia firmy. Przyjmij poziom ufności 0.95.

```r
n <- 200
categories <- c("Bardzo niezadowolony", "Niezadowolony", "Nie ma zdania", "Zadowolony", "Bardzo zadowolony")
x <- c(14, 17, 40, 100, 29)
alpha <- 0.05

binom_ci_exact <- binom.confint(x, n, conf.level = 1 - alpha/5, methods = 'exact')
lower_ci_exact <- binom_ci_exact$lower
upper_ci_exact <- binom_ci_exact$upper

binom_ci_asymptotic <- binom.confint(x, n, conf.level = 1 - alpha/5, methods = 'asymptotic')
lower_ci_asymptotic <- binom_ci_asymptotic$lower
upper_ci_asymptotic <- binom_ci_asymptotic$upper

p <- x/n

df <- data.frame(categories, p, lower_ci_exact, upper_ci_exact, lower_ci_asymptotic, upper_ci_asymptotic)
df$categories <- factor(df$categories, levels = c("Bardzo niezadowolony", "Niezadowolony", "Nie ma zdania", "Zadowolony", "Bardzo zadowolony"))

ggplot(df, aes(x = categories, y = p)) +
  geom_point(color = "black", size = 4) +
  geom_errorbar(aes(ymin = lower_ci_exact, ymax = upper_ci_exact, color = "Dokładny"),
                width = 0.2, alpha = 0.5, linewidth = 1.5) +
  geom_errorbar(aes(ymin = lower_ci_asymptotic, ymax = upper_ci_asymptotic, color = "Asymptotyczny"),
                width = 0.2, linetype = 'dashed', linewidth = 1, alpha = 0.7) +
  labs(title = "Przedział ufności dla prawdopodobieństwa zadowolenia",
       y = "Prawdopodobieństwo",
       x = "Stopień zadowolenia",
       color = "Legend") +
  scale_color_manual(values = c("blue", "deeppink"),
                     labels = c("Dokładny", "Asymptotyczny"),
                     name = "Przedział ufności") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 15),
        axis.text = element_text(size = 10.5),
        axis.title.y = element_text(size = 13),
        axis.title.x = element_text(size = 13),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 10)
  )
```

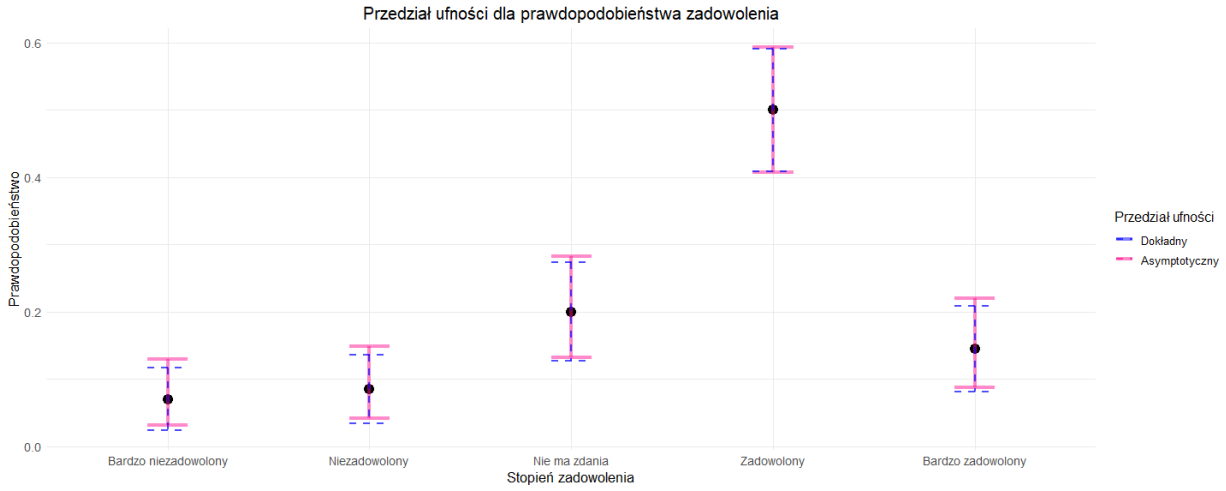
W wyniku działania powyższego kodu otrzymaliśmy przedziały ufności na poziomie ufności 0.95. Zostały one przedstawione w tabelce poniżej.

Stopień zadowolenia	P-stwo	dokładny dolny	dokładny górny	asympto-tyczny dolny	asympto-tyczny górny
Bardzo niezadowolony	0.0700	0.0317	0.1299	0.0235	0.1165
Niezadowolony	0.085	0.0421	0.1487	0.0342	0.1358
Nie ma zdania	0.200	0.1325	0.2822	0.1271	0.2729
Zadowolony	0.500	0.4074	0.5926	0.4089	0.5911
Bardzo zadowolony	0.145	0.0875	0.2201	0.0809	0.2091

Tabela 1: Przedziały ufności dla wektora prawdopodobieństw

W tabeli 1 zostały przedstawione obliczone przedziały ufności dla wektora prawdopodobieństw opisującego stopień zadowolenia z podejścia firmy.

Dodatkowo, sporządziłyśmy wykres, na którym są widoczne dwa rodzaje przedziałów ufności: dokładny (zaznaczony na niebiesko) i asymptotyczny (zaznaczony na różowo). Są one wyznaczane na poziomie ufności 95%, co oznacza, że istnieje 95% pewności, że rzeczywiste prawdopodobieństwo znajduje się w tych przedziałach.



Rysunek 1: Przedziały ufności dla wektora prawdopodobieństw

Na wykresie 1 oraz w tabeli 1 widzimy, że przedziały ufności dla każdej kategorii są stosunkowo wąskie, co wskazuje na precyzję oszacowań prawdopodobieństw. Możemy zauważyć także, że przedziały ufności dla "zadowolony"są najszersze, co może wynikać z największej liczby odpowiedzi w tej kategorii, jednak wciąż pozostają one w akceptowalnym zakresie.

Porównując dokładne przedziały ufności (niebieskie) z asymptotycznymi (różowe) widzimy, że są one zbliżone do siebie, co sugeruje, że obie metody dają podobne wyniki. Górne przedziały są zazwyczaj trochę wyższe w przypadku asymptotycznych przedziałów ufności, dolne są do siebie bardziej zbliżone.

Na podstawie odpowiedzi zaznaczonych przez pracowników oraz wyliczonych przedziałów ufności możemy stwierdzić, że firma dobrze radzi sobie w utrzymaniu równowagi między życiem zawodowym a prywatnym.

4.2 Zadanie 2

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

- chi-kwadrat Pearsona
- chi-kwadrat największej wiarogodności

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

```r
chi2_test <- function(x, p0, n = sum(x), alpha = 0.05, method = c("pearson", "likelihood")) {
  df <- length(x) - 1
  expected <- n*p0

  if (method == "pearson") {
    chi2 <- sum(((x - expected)^2) / expected)
    p_val <- 1 - pchisq(chi2, df = df)
  }

  else if (method == "likelihood") {
    chi2 <- 2 * sum(x * log(x / expected))
    p_val <- 1 - pchisq(chi2, df = df)
  }

  crit_val <- qchisq(1 - alpha, df)

  return(list(chi2 = chi2, p_val = p_val, crit_val = crit_val))
}

set.seed(123)
alpha <- 0.05
n <- 1000
p0 <- c(0.25, 0.50, 0.25)
x <- rmultinom(1, n, p0)

pearson_chi2_test <- chi2_test(x, p0, method = "pearson")
MLE_chi2_test <- chi2_test(x, p0, method = "likelihood")

test_results <- data.frame(
  Test = c("Chi2 Pearson", "Chi2 MLE"),
  Odrzucenie_H0 = c(ifelse(pearson_chi2_test$chi2 > pearson_chi2_test$crit_val, "Tak", "Nie"),
                    ifelse(MLE_chi2_test$chi2 > MLE_chi2_test$crit_val, "Tak", "Nie")),
  Statystyka_chi2 = c(pearson_chi2_test$chi2, MLE_chi2_test$chi2),
  p_wartość = c(pearson_chi2_test$p_val, MLE_chi2_test$p_val),
  Wartość_krytyczna = c(pearson_chi2_test$crit_val, pearson_chi2_test$crit_val)
)

print(test_results)
```

Test	Odrzucenie H_0	Statystyka χ^2	p-wartość	Wartość krytyczna
χ^2 Pearsona	Nie	0.7260	0.6956	5.9915
χ^2 NW	Nie	0.7274	0.6951	5.9915

Tabela 2: Wyniki testów χ^2

Otrzymaliśmy wyniki widoczne w tabeli 2. Hipoteza zerowa w obu testach zakładała, że rzeczywiste prawdopodobieństwo p jest równe p_0. W tabeli widzimy, że zarówno w przypadku test χ^2 Pearsona, jak i w teście χ^2 największej wiarygodności nie została ona odrzucona. Oznacza to, że nie ma wystarczających dowodów na to, że rzeczywiste prawdopodobieństwo p różni się od zakładanego p_0.

Wynika to z faktu, że wartości statystyk χ^2 dla obu testów (równe odpowiednio 0.7260 i 0.7274) są znacznie mniejsze od wartości krytycznej równej 5.9915. Oznacza to, że obserwowane dane nie dostarczają dowodów na odrzucenie hipotezy zerowej.

4.3 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.

```r
alpha <- 0.05
data_DK <- subset(data, DZIAŁ == "DK")

table_DK <- table(data_DK$PYT_1)
p0 <- rep(1/5, 5)

pearson_chi2_test <- chi2_test(table_DK, p0, method = "pearson")
MLE_chi2_test <- chi2_test(table_DK, p0, method = "likelihood")

test_results <- data.frame(
  Test = c("Chi2 Pearson", "Chi2 MLE"),
  Odrzucenie_H0 = c(ifelse(pearson_chi2_test$chi2 > pearson_chi2_test$
    crit_val, "Tak", "Nie"),
                    ifelse(MLE_chi2_test$chi2 > MLE_chi2_test$crit_val, "
    Tak", "Nie")),
  Statystyka_chi2 = c(pearson_chi2_test$chi2, MLE_chi2_test$chi2),
  p_wartość = c(pearson_chi2_test$p_val, MLE_chi2_test$p_val),
  Wartość_krytyczna = c(pearson_chi2_test$crit_val, pearson_chi2_test$
    crit_val)
)

print(test_results)
```

W wyniku działania powyższego kodu otrzymaliśmy wyniki widoczne w tabeli poniżej.

Test	Odrzucenie H_0	Statystyka χ^2	p-wartość	Wartość krytyczna
χ^2 Pearsona	Tak	64.8571	$2.7578 * 10^{-13}$	9.4877
χ^2 NW	Tak	52.5271	$1.0702 * 10^{-10}$	9.4877

Tabela 3: Wyniki testów χ^2

Hipoteza zerowa mówi nam o tym, że rozkład odpowiedzi jest równomierny, czyli każda z pięciu możliwych odpowiedzi ma takie samo prawdopodobieństwo. W obu testach została ona odrzucona. Oznacza to, że istnieją statystycznie istotne dowody na to, że rozkład odpowiedzi nie jest równomierny.

Potwierdzają to wysokie wartości statystyki χ^2 równe odpowiednio 64.8571 (test χ^2 Pearsona) i 52.5271 (test χ^2 największej wiarygodności), które są znacznie wyższe od wartości krytycznej równej 9.4877.

Analizując p-wartości dla obu testów ($2.7578 * 10^{-13}$ i $1.0702 * 10^{-10}$) widzimy, że są one znacznie niższe od poziomu istotności równego 0.05. To także prowadzi do odrzucenia hipotezy zerowej.

Podsumowując, z wyników testów χ^2 Pearsona i największej wiarogodności widocznych w tabeli 3 wynika, że rozkład odpowiedzi wśród pracowników Działu Kreatywnego na pytanie dotyczące podejścia firmy do utrzymania równowagi między życiem zawodowym a prywatnym nie jest równomierny. Istnieją silne dowody statystyczne na to, że różne odpowiedzi mają różne prawdopodobieństwa.

5 Część II

5.1 Zadanie 4

Zapoznaj się z funkcją *fisher.test* z pakietu *stats*.

```
library(stats)
help(fisher.test)
```

Funkcja *fisher.test* z pakietu *stats* służy do przeprowadzania dokładnego testu Fishera. Jest ona używana głównie do zbadania czy istnieje statystycznie istotna zależność między dwiema zmiennymi kategorycznymi. Hipoteza zerowa dokładnego testu Fishera mówi nam o tym, że istnieje niezależność między zmiennymi. Funkcja *fisher.test* jest szczególnie użyteczna, gdy liczności w tabeli są małe, co może uczynić test χ^2 mniej wiarygodnym.

5.2 Zadanie 5

Korzystajac 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 prawdopodobieństwu tego, że na stanowisku kierowniczym pracuje mężczyzna?

```
alpha = 0.05

fisher_test <- fisher.test(data$PŁEĆ, data$CZY_KIER, conf.level = 1 -
    alpha)$p.value
cat("PŁEC a CZY_KIER - ", ifelse(fisher_test > alpha, "Nie zależy", "Zale
    ży"), ", bo p-value =", fisher_test, "\n")
```

Hipoteza zerowa testu Fishera w tym przypadku mówi nam o tym, że zmienna **PŁEĆ** jest niezależna od zmiennej **CZY_KIER**, czyli że prawdopodobieństwo, że osoba pracuje na stanowisku kierowniczym, jest takie samo dla mężczyzn i kobiet.

W wyniku działania powyższego kodu otrzymaliśmy p-wartość równą 0.6659029, jest ona większa niż poziom istotności równy 0.05, co pozwala nam wnioskować, że nie istnieją statystycznie istotne dowody na to, że możemy odrzucić hipotezę zerową, czyli że zmienna **PŁEĆ** jest niezależna od zmiennej **CZY_KIER**. Innymi słowy, prawdopodobieństwo, że na stanowisku kierowniczym pracuje kobieta jest równe prawdopodobieństwu, że na tym stanowisku pracuje mężczyzna.

5.3 Zadanie 6

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

- zajmowanie stanowiska kierowniczego nie zależy od wieku (**CZY_KIER** oraz **WIEK_KAT**),
- zajmowanie stanowiska kierowniczego nie zależy od staż pracy (**CZY_KIER** oraz **STAŻ**),
- zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od zajmowanego stanowiska (**PYT_2** oraz **CZY_KIER**),
- zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od stażu (**PYT_2** oraz **STAŻ**),
- zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od płci (**PYT_2** oraz **PŁEĆ**),
- zadowolenie z wynagrodzenia w pierwszym badanym okresie nie zależy od wieku (**PYT_2** oraz **WIEK_KAT**).

```
alpha = 0.05
```

Aby zweryfikować hipotezę dotyczącą zależności między odpowiednimi dwoma zmiennymi na poziomie istotności 0.05, używamy testu Freemana-Haltona, który jest rozszerzeniem dokładnego testu Fishera do większych tabel. Hipoteza zerowa tego testu zakłada, że dwie zmienne są niezależne.

```
fisher_test_1 <- fisher.test(data$CZY_KIER, data$WIEK_KAT, conf.level = 1
    - alpha)$p.value
cat("CZY_KIER a WIEK_KAT - ", ifelse(fisher_test_1 > alpha, "Nie zależy",
    "Zależy"), ", bo p-value =", fisher_test_1, "\n")
```

W wyniku działania powyższego kodu widzimy, że p-wartość równa 0.7823002 jest znacznie większa od poziomu istotności równego 0.05, czyli nie mamy podstaw do odrzucenia hipotezy zerowej o niezależności zmiennych **CZY_KIER** i **WIEK_KAT**. Możemy stwierdzić, że wiek nie ma statystycznie istotnego wpływu na to, czy ktoś zajmuje stanowisko kierownicze.

```
fisher_test_2 <- fisher.test(data$CZY_KIER, data$STAŻ, conf.level = 1 -
    alpha)$p.value
cat("CZY_KIER a STAZ - ", ifelse(fisher_test_2 > alpha, "Nie zależy", "
    Zależy"), ", bo p-value =", fisher_test_2, "\n")
```

W wyniku działania powyższego kodu widzimy, że p-wartość równa $6.537896 * 10^{-5}$ jest znacznie mniejsza od poziomu istotności równego 0.05, czyli mamy podstawy do odrzucenia hipotezy zerowej. Oznacza to, że zmienne **CZY_KIER** i **STAŻ** nie są niezależne. Możemy stwierdzić, że staż pracy ma statystycznie istotny wpływu na zajmowanie stanowiska kierowniczego.

```
fisher_test_3 <- fisher.test(data$PYT_2, data$CZY_KIER, conf.level = 1 -
    alpha)$p.value
cat("PYT_2 a CZY_KIER - ", ifelse(fisher_test_3 > alpha, "Nie zależy", "
    Zależy"), ", bo p-value =", fisher_test_3, "\n")
```

W wyniku działania powyższego kodu widzimy, że p-wartość równa 0.04429734 jest mniejsza od poziomu istotności równego 0.05, czyli mamy wystarczające podstawy do odrzucenia hipotezy zerowej. Oznacza to, że zmienne **PYT_2** i **CZY_KIER** nie są niezależne. Możemy stwierdzić, że poziom zadowolenia z wynagrodzenia różni się między osobami zajmującymi stanowiska kierownicze a innymi pracownikami.

```
fisher_test_4 <- fisher.test(data$PYT_2, data$STAŻ, conf.level = 1 -
    alpha)$p.value
cat("PYT_2 a STAZ - ", ifelse(fisher_test_4 > alpha, "Nie zależy", "Zależ
    y"), ", bo p-value =", fisher_test_4, "\n")
```

W wyniku działania powyższego kodu widzimy, że p-wartość równa 0.01069023 jest mniejsza od poziomu istotności równego 0.05, czyli mamy podstawy do odrzucenia hipotezy zerowej. Oznacza to, że zmienne **PYT_2** i **STAŻ** nie są niezależne. Możemy stwierdzić, że poziom zadowolenia z wynagrodzenia różni się w zależności od długości stażu prazy.

```
fisher_test_5 <- fisher.test(data$PYT_2, data$PŁEĆ, conf.level = 1 -
    alpha)$p.value
cat("PYT_2 a PLEC - ", ifelse(fisher_test_5 > alpha, "Nie zależy", "Zależ
    y"), ", bo p-value =", fisher_test_5, "\n")
```

W wyniku działania powyższego kodu widzimy, że p-wartość równa 0.4758086 jest większa od poziomu istotności równego 0.05, czyli nie mamy podstaw do odrzucenia hipotezy zerowej. Oznacza to, że zmienne **PYT_2** i **PŁEĆ** są niezależne. Możemy stwierdzić, że płeć nie ma wpływu na poziom zadowolenia z wynagrodzenia.

```
fisher_test_6 <- fisher.test(data$PYT_2, data$WIEK_KAT, conf.level = 1 -
    alpha, workspace= 300000)$p.value
cat("PYT_2 a WIEK_KAT - ", ifelse(fisher_test_6 > alpha, "Nie zależy", "
    Zależy"), ", bo p-value =", fisher_test_6, "\n")
```

W wyniku działania powyższego kodu widzimy, że p-wartość równa 0.319352 jest znacznie większa od poziomu istotności równego 0.05, czyli nie mamy podstaw do odrzucenia hipotezy zerowej. Oznacza to, że zmienne **PYT_2** i **WIEK_KAT** są niezależne. W tym przypadku możemy stwierdzić, że możemy stwierdzić, że wiek nie ma statystycznie istotnego wpływu na poziom zadowolenia z wynagrodzenia.

6 Część III

6.1 Zadanie 7

Zapoznaj się z funkcją *chisq.test* z pakietu *stats* oraz *assocplot* z pakietu *graphics*.

```
library(stats)
help(chisq.test)

```

```r
library(graphics)
help(assocplot)
```

Funkcja *chisq.test* z pakietu *stats* wykonuje test χ^2 niezależności dla tabel kontyngencji. Test ten jest używany do sprawdzenia, czy istnieje zależność między dwiema zmiennymi kategorycznymi.

Działanie funkcji *chisq.test* polega na porównaniu rzeczywistych wyników z tymi, których możnaby oczekiwać, gdyby zmienne były niezależne. W wyniku testu otrzymujemy wartość statystyki χ^2 oraz p-wartość, która informuje nas o istotności wyników. Im mniejsza p-wartość, tym większe prawdopodobieństwo, że istnieje istotna zależność między zmiennymi.

Natomiast funkcja *assocplot* z pakietu *graphics* wizualizuje strukturę zależności między zmiennymi kategorycznymi na podstawie wyników testu χ^2. Wizualizacja ta może przyjąć formę wykresu heatmap, na którym są przedstawione wartości statystyki χ^2 lub wartości p-wartości dla każdej komórki tabeli.

6.2 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.

```r
alpha <- c(0.01, 0.05)
table <- table(data$PYT_2, data$CZY_KIER)

test_chi2_p_val_correct <- chisq.test(table, correct=TRUE)$p.val # z poprawką
test_chi2_p_val_no_correct <- chisq.test(table, correct=FALSE)$p.val # bez poprawki

for (i in 1:length(alpha)) {
  cat(" alpha = ", alpha[i], ", test chi2 z poprawką - ", ifelse(test_chi2_p_val_correct > alpha[i], "Nie zależy", "Zależy"), ", bo p-value =", test_chi2_p_val_correct, "\n")
  cat(" alpha = ", alpha[i], ", test chi2 bez poprawki - ", ifelse(test_chi2_p_val_no_correct > alpha[i], "Nie zależy", "Zależy"), ", bo p-value =", test_chi2_p_val_no_correct, "\n")


  # z zadania 6 (do porównania)
  fisher_test_3 <- fisher.test(table, conf.level = 1 - alpha[i])$p.value
  cat(" alpha = ", alpha[i], ", fisher test - ", ifelse(fisher_test_3 > alpha[i], "Nie zależy", "Zależy"), ", bo p-value =", fisher_test_3, "\n\n")

}
```

Wyniki testu χ^2 oraz testu Fishera są używane do oceny zależności między dwiema zmiennymi kategorycznymi. W tym przypadku badamy zależność pomiędzy zajmowanym stanowiskiem, a zadowoleniem z wynagrodzenia w pierwszym badanym okresie.

Dla $\alpha = 0.01$ otrzymaliśmy, że test χ^2 z poprawką oraz bez poprawki wskazuje, że są podstawy do odrzucenia niezależności między zadowoleniem z wynagrodzenia, a zajmowanym stanowiskiem, ponieważ obie p-wartości (0.004397081 dla testu z poprawką i 0.004397081 dla testu bez poprawki) są mniejsze niż poziom istotności $\alpha = 0.01$. Test

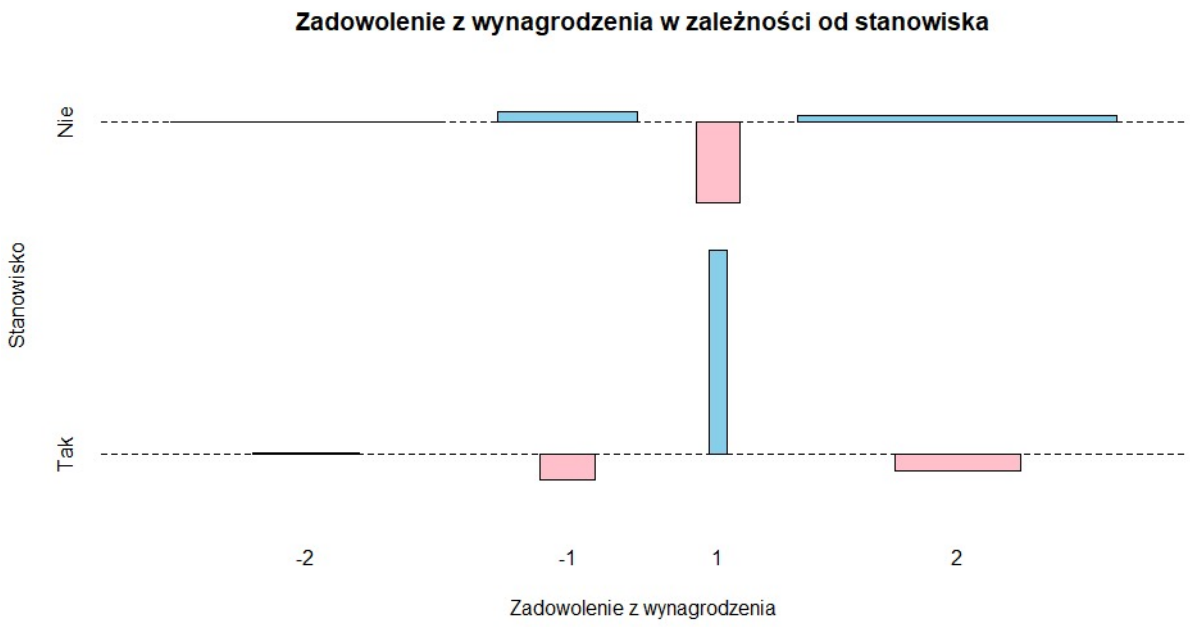
Fishera wskazuje natomiast, że nie ma podstaw do odrzucenia niezależności zmiennej opisującej zadowolenie z wynagrodzenia i zajmowanego stanowiska, ponieważ p-wartość równa 0.04429734 jest większa niż α równe 0.01.

Dla $\alpha = 0.05$ zarówno test χ^2 z poprawką, jak i bez poprawki, wskazuje, że mamy podstawy do odrzucenia niezależności między zadowoleniem z wynagrodzenia, a zajmowanym stanowiskiem, ponieważ p-wartości są mniejsze od poziomu istotności alfa równego 0.05. Test Fishera prowadzi do analogicznych wniosków choć p-wartość jest nieco większa (ale wciąż mniejsza od poziomu istotności) niż dla testów χ^2.

Podsumowując, na poziomie istotności 0.01 wyniki testu Fishera nie są zgodne z wynikami testu χ^2, co może sugerować, że test Fishera jest bardziej konserwatywny przy mniejszych poziomach istotności. Na poziomie istotności 0.05, oba testy zgodnie wskazują, że mamy podstawy do odrzucenia niezależności między zmiennymi.

```
assocplot(table, main="Zadowolenie z wynagrodzenia w zależności od
    stanowiska",
          xlab="Zadowolenie z wynagrodzenia", ylab="Stanowisko",
          col=c("skyblue", "pink"))
```

Korzystając z kodu widocznego powyżej, który wykorzystuje funkcję *assocplot* zwizualizowałyśmy wyniki na wykresie, który jest widoczny poniżej.



Rysunek 2: Wyniki testu *chisq.test*

Jak możemy zauważyć na rysunku 2, dla osób, które nie zajmują kierowniczego stanowiska przy zadowoleniu z wynagrodzenia na poziomie 1 i 2, widzimy niebieskie prostokąty, co sugeruje, że osoby na stanowiskach nie kierowniczych częściej są zadowolone z wynagrodzenia niż by to wynikało z oczekiwań (nadmiar obserwacji). Przy zadowoleniu na poziomie -1, widzimy różowe prostokąt, co sugeruje, że mniej osób na stanowiskach nie kierowniczych jest niezadowolonych z wynagrodzenia niż się spodziewano (niedobór obserwacji). Natomiast dla osób zajmujących kierownicze, przy zadowoleniu z wynagrodzenia na poziomie -1 i 2, widzimy różowe prostokąty, co sugeruje, że osoby na stanowiskach kierowniczych rzadziej są niezadowolone lub bardzo zadowolone z wynagrodzenia niż by to wynikało z

oczekiwań (niedobór obserwacji). Przy zadowoleniu na poziomie 1, widzimy niebieski prostokąt, co sugeruje, że więcej osób na stanowiskach kierowniczych jest zadowolonych z wynagrodzenia niż się spodziewano (nadmiar obserwacji).

Podsumowując, ogólne wnioski, które możemy wyciągnąć na podstawie analizy wykresu 2 mówią nam o tym, że osoby na stanowiskach nie kierowniczych częściej są zadowolone z wynagrodzenia (poziom 1 i 2) niż osoby na stanowiskach kierowniczych. Z kolei osoby na stanowiskach kierowniczych rzadziej są bardzo zadowolone lub niezadowolone z wynagrodzenia (poziom 2 i -1) w porównaniu do oczekiwań. Potwierdza nam to wyniki z pierwszej części zadania.

6.3 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 22, 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 $n = 1000$.

```r
# zapoznanie
help(rmultinom)
```

Funkcja *rmultinom* z pakietu *stats* służy do generowania losowych próbek z rozkładu wielomianowego. Przyjmuje ona jako argumenty wartość n odpowiadającą liczbie generowanych próbek, parametry "*size* opisujący liczbę prób w każdej próbce oraz "*prob*", czyli wektor prawdopodobieństw.

W przypadku $n = 50$ test χ^2 nie działał, dlatego dla ulepszenia kodu do próbki generowanej z pomocą funkcji *rmultinom* dodajemy wartość 0.00000001.

```r
set.seed(123)

p <- c(1/40, 3/40, 19/40, 17/40)

alpha <- c(0.01, 0.05)

test_power <- function(p, n, alpha, N = 500) {
  fisher_count <- 0
  chi2_count_correct <- 0
  chi2_count_no_correct <- 0

  for (i in 1:N) {
    X <- rmultinom(1, n, p)

    while (all(X == 0)) {
      X <- rmultinom(1, n, p)
    }

    fisher_p_val <- fisher.test(matrix(X, nrow = 2))$p.val
    chi2_p_val_correct <- chisq.test(matrix(X+0.00000001, nrow = 2),
    correct = TRUE)$p.val
    chi2_p_val_no_correct <- chisq.test(matrix(X+0.00000001, nrow = 2),
    correct = FALSE)$p.val

    if (fisher_p_val < alpha) {
      fisher_count <- fisher_count + 1
    }
    if (chi2_p_val_correct < alpha) {
      chi2_count_correct <- chi2_count_correct + 1
```

```
    }
    if (chi2_p_val_no_correct < alpha) {
      chi2_count_no_correct <- chi2_count_no_correct + 1
    }
  }

  return(c(fisher_count / N, chi2_count_correct / N, chi2_count_no_
   correct / N))
}

for (i in 1:length(alpha)) {
  results <- data.frame(
    test = c("Fisher", "Chi-2 z poprawką", "Chi-2 bez poprawki"),
    power_n_50 = test_power(p, 50, alpha[i]),
    power_n_100 = test_power(p, 100, alpha[i]),
    power_n_1000 = test_power(p, 1000, alpha[i])
  )

  print(results)
  cat("\n\n")

  table <- xtable(results)
  print(table)
  cat("\n\n")
}
```

Wyniki działania powyższego kodu zostały przedstawione w tabeli

test	n = 50	n = 100	n = 1000
Fisher	0.040	0.120	1.000
χ^2 z poprawką	0.014	0.092	1.000
χ^2 bez poprawki	0.056	0.148	1.000

Tabela 4: Moce testów Fischera oraz χ^2 dla różnych n i $\alpha = 0.01$

test	n = 50	n = 100	n = 1000
Fisher	0.124	0.304	1.000
χ^2 z poprawką	0.070	0.238	1.000
χ^2 bez poprawki	0.186	0.378	1.000

Tabela 5: Moce testów Fischera oraz χ^2 dla różnych n i $\alpha = 0.05$

W tabeli 4 zostały przedstawione moce testów Fischera oraz χ^2 Pearsona z poprawką i bez dla $n \in \{50, 100, 1000\}$ oraz $\alpha = 0.01$. W tabeli 5 widoczne są analogiczne wyniki tylko dla poziomu istotności $\alpha = 0.05$.

Jak możemy zauważyć, w obu przypadkach test Fishera zdaje się być bardziej mocny (czyli lepiej wykrywa różnice) niż test χ^2 Pearsona z poprawką dla naszych danych oraz rozmiarów próbek. Oba są jednak mniej mocne od testu χ^2 bez poprawki. Analogicznie test Fishera wydaje się być nieco mniej skuteczny niż test χ^2 bez poprawki i bardziej skuteczny niż test χ^2 z poprawką, zwłaszcza dla mniejszych próbek. Widzimy także, że moc testów rośnie wraz ze wzrostem rozmiaru próbki, co sugeruje, że większe próbki są bardziej skuteczne w wykrywaniu różnic. Porównując wyniki dla różnych poziomów istotności możemy zauważyć, że czym dla wyższego poziomu istotności i tego samego n moc testu jest większa. Wnioski te są zgodne z naszą intuicją.

6.4 Zadanie 10

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

```
alpha <- c(0.01, 0.05)
table <- table(data$PYT_2, data$CZY_KIER)

# dwudzielcza
table <- rbind(table, colSums(table))
table <- cbind(table, rowSums(table))
print(table)
```

Otrzymaliśmy tablicę dwudzielczą widoczną poniżej

	Nie	Tak	Suma
-2	64	10	74
-1	18	2	20
1	0	2	2
2	91	13	104
Suma	173	27	200

Tabela 6: Tablica dwudzielcza

```
likehood_test <- function(x) {
  R <- dim(x)[1]
  C <- dim(x)[2]

  df <- (R - 2)*(C - 2) # stopnie swobody

  n <- x[R, C]

  lambda_part <- matrix(0, nrow = R - 1 , ncol = C - 1)
  for (i in 1:(R - 1)) {
    for (j in 1:(C - 1)) {
      n_i_plus <- x[i, C]
      n_plus_j <- x[R, j]
      n_i_j <- x[i, j]

      lambda_part[i, j] <-  ((n_i_plus * n_plus_j) / (n * n_i_j)) ^ n_i_j
    }
  }
  lambda <- prod(lambda_part)
  G2 <- -2 * log(lambda)

  p_val <- 1 - pchisq(G2, df = df)

  return(data.frame(G2 = G2, p_val = p_val))
}

cat("\nWyniki:")
for (i in 1:length(alpha)) {
  cat('alpha =', alpha[i], ':', ifelse(result[1, 2] > alpha[i], "
   Niezalezne", "Zalezne"), ", p-wartość = ", result[1, 2], '\n')
}
```

G2	p-wartość
8.3285	0.0397

Tabela 7: Wartości statystyki G2 oraz p-wartość

Wartości statystyki G2 oraz p-wartość zostały umieszczone w tabeli 7. Jak możemy zauważyć, p-wartość (0.0397) jest większa od poziomu istotności 0.01, ale mniejsza od 0.05. Oznacza to, że na poziomie istotności 0.01 mamy wystarczająco dowodów, aby odrzucić hipotezę zerową o niezależności zmiennych, ale na poziomie 0.05 nie mamy wystarczająco silnych dowodów, aby to zrobić. Oznacza to, że dla poziomu istotności 0.01 możemy wnioskować, że zmienne są niezależne, a dla poziomu istotności 0.05, że nie są.

7 Część IV i V

7.1 Zadanie 11

Przeprowadzone wśród brytyjskich mężczyzn badanie trwające 20 lat wykazało, że odsetek zmarłych (na rok) z powodu raka płuc wynosił $0,00140$ wśród osób palących papierosy i $0,00010$ wśród 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?

```r
deaths_data <- matrix(c(0.00140, 0.00010, 0.00669, 0.00413), nrow=2,
    byrow=TRUE)
colnames(deaths_data) <- c("Palący", "Niepalący")
rownames(deaths_data) <- c("Rak płuc", "Choroba serca")
print(deaths_data)

diff_proportion <- deaths_data[, "Palący"] - deaths_data[, "Niepalący"]

RR <- deaths_data[, "Palący"] / deaths_data[, "Niepalący"]

chance_1 <- deaths_data[, "Palący"] / (1 - deaths_data[, "Palący"])
chance_2<- deaths_data[, "Niepalący"] / (1 - deaths_data[, "Niepalący"])
OR <- chance_1 / chance_2

result <- data.frame(
  Różnica_proporcji = diff_proportion,
  RR = RR,
  OR = OR
)

print(result)
```

W poniższej tabeli została przedstawiona różnica proporcji (RP), ryzyko względne RR oraz iloraz szans OR, które zostały przez nas wykorzystane do analizy związku między paleniem papierosów, a śmiercią z powodu raka płuc i choroby serca.

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

Tabela 8: Związek pomiędzy paleniem papierosów, a śmiercią z powodu danych chorób

Analizując wyniki widoczne w tabeli 8 możemy zauważyć, że:

- Dla raka płuc:
  - Różnica proporcji (RP) wynosi 0.0013, co oznacza, że odsetek zmarłych na raka płuc wśród palących jest o 0.0013 wyższy niż wśród niepalących.
  - Ryzyko względne (RR) wynosi 14, co oznacza, że osoby palące mają 14-krotnie większe ryzyko zgonu na raka płuc w porównaniu z osobami niepalącymi.
  - Iloraz szans (OR) wynosi 14.0182, co oznacza, że osoby palące mają ponad 14-krotnie większe szanse na zgon z powodu raka płuc w porównaniu z osobami niepalącymi.
- Dla chorób serca:
  - Różnica proporcji (RP) wynosi 0.0026, co oznacza, że odsetek zmarłych na chorobę serca wśród palących jest o 0.0026 wyższy niż wśród niepalących.
  - Ryzyko względne (RR) wynosi 1.6199, co oznacza, że osoby palące mają około 1.62-krotnie większe ryzyko zgonu na chorobę serca w porównaniu z osobami niepalącymi.
  - Iloraz szans (OR) wynosi 1.6240, co oznacza, że osoby palące mają około 1.62-krotnie większe szanse na zgon z powodu choroby serca w porównaniu z osobami niepalącymi.

Jak możemy zauważyć, związek między paleniem papierosów, a śmiercią z powodu raka płuc jest znacznie silniejszy niż związek między paleniem a śmiercią z powodu choroby serca. Obserwację tą potwierdza zarówno różnica proporcji, jak i ryzyko względne, które są znacznie większe w przypadku raka płuc. Analiza ilorazu szans prowadzi do podobnych wniosków.

7.2 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.

Tabela 1: Wyniki dotyczące śmiertelności w wypadkach samochodowych na Florydzie w 2008 roku.

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

a) 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.

b) 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.

c) 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?

```r
#do wszystkich podpunktów
car_crash <- matrix(c(1085, 55623, 703, 441239), nrow=2, byrow=TRUE)
colnames(car_crash) <- c("Śmiertelny", "Nieśmiertelny")
rownames(car_crash) <- c("Bez pasów", "Z pasami")
print(car_crash)
```

```r
# a)
prob_death_no_seatbelt <- car_crash["Bez pasów", "Śmiertelny"] / sum(car_
    crash["Bez pasów", ])

prob_death_seatbelt <- car_crash["Z pasami", "Śmiertelny"] / sum(car_
    crash["Z pasami", ])

# Wyniki
cat("Warunkowe prawdopodobieństwo śmierci bez pasów:", prob_death_no_
    seatbelt, "\n")
cat("Warunkowe prawdopodobieństwo śmierci z pasami:", prob_death_seatbelt
    , "\n")
```

W wyniku działania powyższego kodu otrzymujemy, że warunkowe prawdopodobieństwo śmierci bez pasów wynosi około 0.0191, natomiast z użyciem pasów bezpieczeństwa jest to około 0.0016. Pozwala nam to wnioskować, że osoby używające pasów bezpieczeństwa mają znacznie niższe warunkowe prawdopodobieństwo śmierci w wypadkach samochodowych niż osoby, które nie ich nie używają. Potwierdza to naszą intuicję, że pasy bezpieczeństwa znacząco zmniejszają ryzyko poważnych obrażeń lub śmierci w wypadkach samochodowych.

```r
# b)
prob_seatbelt_fatal <- car_crash["Z pasami", "Śmiertelny"] / sum(car_
    crash[ , "Śmiertelny"])

prob_seatbelt_nonfatal <- car_crash["Z pasami", "Nieśmiertelny"] / sum(
    car_crash[ , "Nieśmiertelny"])

# Wyniki
cat("Warunkowe prawdopodobieństwo użycia pasa bezpieczeństwa w wypadkach
    śmiertelnych:", prob_seatbelt_fatal, "\n")
cat("Warunkowe prawdopodobieństwo użycia pasa bezpieczeństwa w wypadkach
    nieśmiertelnych:", prob_seatbelt_nonfatal, "\n")
```

W wyniku działania powyższego kodu otrzymaliśmy, że warunkowe prawdopodobieństwo użycia pasa bezpieczeństwa w wypadkach śmiertelnych to około 0.3932, natomiast w wypadkach nieśmiertelnych wynosi ono około 0.8881. Pozwala nam to wnioskować, że osoby, które przeżyły wypadki samochodowe częściej korzystały z pasa bezpieczeństwa w porównaniu z osobami, które w nich zginęły. To sugeruje, że używanie pasa bezpieczeństwa może

znacząco zmniejszyć ryzyko śmierci w wypadku i tym samym potwierdza wniosek płynący z podpunktu a, mówiący o tym, że zapięte pasy znacznie zwiększają nasze bezpieczeństwo.

c) Wybór zmiennej objaśnianej: Najbardziej naturalnym wyborem zmiennej objaśnianej wydaje się być śmiertelność w wypadkach samochodowych, ponieważ to właśnie to zjawisko jest zwykle badane w kontekście bezpieczeństwa w ruchu drogowym. W związku z tym bardziej naturalne będzie podejście opisane w podpunkcie a).

```r
# zmienna objasniania - czy wypadek był śmiertelny
diff_proportion_a <-  prob_death_no_seatbelt - prob_death_seatbelt

RR_a <-  prob_death_no_seatbelt / prob_death_seatbelt

chance_1_a <- prob_death_no_seatbelt  / (1 - prob_death_no_seatbelt )
chance_2_a <- prob_death_seatbelt  / (1 - prob_death_seatbelt)
OR_a <- chance_1_a / chance_2_a

result <- data.frame(
  Ryzyko_śmierci = c("Bez pasów vs Z pasami"),
  Różnica_proporcji = c(diff_proportion_a),
  RR = c(RR_a),
  OR = c(OR_a)
)
print(result)
```

Ryzyko śmierci	Różnica proporcji	RR	OR
Bez pasów vs Z pasami	0.0175	12.0281	12.2432

Tabela 9: Ryzyko śmierci w wypadkach samochodowych

Jak możemy zauważyć w tabeli 9, różnica proporcji wynosi około 0.0175, co oznacza, że osoby bez pasów mają o 1.75% wyższe prawdopodobieństwo śmierci w wypadku w porównaniu do osób z zapiętymi pasami bezpieczeństwa. Ryzyko względne wynosi około 12.03, co oznacza, że osoby bez pasów mają ponad 12 razy większe ryzyko śmierci w wypadku w porównaniu do osób z pasami. Iloraz szans wynosi około 12.24, co oznacza, że szansa na śmierć w wypadku dla osób bez pasów jest ponad 12 razy większa niż dla osób z pasami.

Jak możemy zauważyć, ryzyko względne (RR) i iloraz szans (OR) są zbliżone. Może to wynikać z faktu, że śmiertelność w wypadkach samochodowych jest rzadkim zdarzeniem - w podpunkcie a) obliczyłyśmy, że dla wypadków bez użycia pasów bezpieczeństwa wynosi około 0.0191, natomiast z ich użyciem jest to około 0.0016. Jak widzimy, obie wartości są dość niewielkie, mniejsze od 0.02. Gdy zdarzenie jest rzadkie (niska częstość występowania), różnice między RR i OR są niewielkie. W takich sytuacjach te dwie miary są często używane zamiennie, ponieważ dają podobne wartości.

7.3 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.

```r
library(DescTools)
help("GoodmanKruskalTau")
help("GoodmanKruskalGamma")
```

W tym zadaniu skorzystałyśmy z funkcji *GoodmanKruskalTau* i *GoodmanKruskalGamma* z biblioteki *DescTools*. Służą one do obliczenia współczynników τ i γ, które mierzą siłę zależności między zmiennymi. Współczynnik τ używamy, gdy mamy do czynienia ze zmiennymi nominalnymi, czyli takimi, które nie mają naturalnego porządku (np. odpowiedzi "TAK"i "NIE"). Przyjmuje on wartości $0 \leq \tau \leq 1$. τ równa 0 oznacza, że zmienne są niezależne. Współczynnika γ używamy natomiast, gdy mamy do czynienia ze zmiennymi porządkowymi, czyli takimi, które mają swój naturalny porządek (np. liczby w skali 1-5). Przyjmuje on wartości z przedziału $-1 \leq \gamma \leq 1$. Podobnie jak w przypadku τ, γ równa 0 oznacza, że zmienne są niezależne.

W pierwszym podpunkcie mamy do czynienia ze zmiennymi nominalnymi, czyli takimi, które nie mają naturalnego porządku - zmienne **CZY_ZADOW** oraz **CZY_KIER** przyjmują wartości "TAK"lub "NIE". W związku z tym skorzystamy ze współczynnika τ.

```r
#zadowolenie z wynagrodzenia w pierwszym badanym okresie i zajmowane
    stanowisko
table_1 <- table(data$CZY_ZADOW, data$CZY_KIER)
GoodmanKruskalTau(table_1, direction = "row")
GoodmanKruskalTau(table_1, direction = "column")
```

W wyniku działania funkcji otrzymaliśmy wartość τ równą około $0.000409 \simeq 0$. Na tej podstawie możemy wnioskować, że zmienne te są prawie niezależne.

W drugim podpunkcie mamy do czynienia ze zmiennymi porządkowymi, czyli takimi, które mają natutalny porządek - zmienna **PYT_2** przyjmuje wartości $-2, -1, 1, 2$, zmienna **STAŻ** przyjmuje wartości $1, 2, 3$. W związku z tym skorzystamy ze współczynnika γ.

```r
#zadowolenie z wynagrodzenia w pierwszym badanym okresie i staż pracy
table_2 <- table(data$PYT_2, data$STAŻ)
GoodmanKruskalGamma(table_2)
```

W wyniku działania funkcji otrzymaliśmy wartość γ równą około 0.090843. Na tej podstawie możemy wnioskować, że między zmiennymi występuje niewielka siła zależności.

W trzecim podpunkcie mamy do czynienia z jedną zmienną nominalną i z jedną porządkową - zmienna **CZY_KIER** przyjmuje wartości "TAK"i "NIE", natomiast zmienna **STAŻ** przyjmuje wartości $1, 2, 3$. W związku z tym sprowadziłyśmy zmienną **STAŻ** do postaci nominalnej. W tym celu stworzyłyśmy pomocniczą zmienną **CZY_STAŻ_DŁUGI**, która przyjmuje wartość "Nie", gdy pracownik pracuje w firmie poniżej 3 lat oraz wartość "Tak", gdy powyżej. Po tej poprawce mamy 2 zmienne nominalne, więc możemy skorzystać ze współczynnika τ.

```r
#zajmowane stanowisko i staż pracy
data$STAŻ_CZY_DŁUGI <- ifelse(data$STAŻ <= 2, "Nie", "Tak")
table_3 <- table(data$CZY_KIER, data$STAŻ_CZY_DŁUGI)
GoodmanKruskalTau(table_3, direction = "row")
GoodmanKruskalTau(table_3, direction = "column")
```

W wyniku działania funkcji otrzymaliśmy wartość τ równą około 0.103113. Na tej podstawie możemy wnioskować, że między zmiennymi występuje niewielka siła zależności.

7.4 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.

```
table <- table(data$PYT_2, data$STAŻ)
table <- rbind(table, colSums(table))
table <- cbind(table, rowSums(table))
print(table)
```

	1	2	3	Suma
-2	20	45	9	74
-1	3	17	0	20
1	0	0	2	2
2	18	78	8	104
Suma	41.00	140.00	19.00	200.00

Tabela 10: Tablica dwudzielcza z danymi do zadania 14

```
correspondence_analysis <- function(x, plot_type = "principal-principal")
    {
  nrows <- dim(x)[1]
  ncols <- dim(x)[2]

  n <- x[nrows, ncols]

  p <- x / n
  P <- p[1:(nrows - 1), 1:(ncols - 1)]

  r <- p[1:(nrows - 1), ncols]
  c <- p[nrows, 1:(ncols - 1)]

  D_r <- diag(r)
  D_c <- diag(c)

  R <- solve(D_r) %*% P
  C <- P %*% solve(D_c)


  A <- solve(sqrt(D_r)) %*% (P - r %*% t(c)) %*% solve(sqrt(D_c))
  svd_A <- svd(A)
  U <- svd_A$u
  V <- svd_A$v
  Gamma <- svd_A$d

  F <- solve(sqrt(D_r)) %*% U %*% diag(Gamma)
  G <- solve(sqrt(D_c)) %*% V %*% diag(Gamma)

  F_std <- solve(sqrt(D_r)) %*% U
  G_std <- solve(sqrt(D_c)) %*% V

  lambda <- sum(Gamma^2)
```

```
33
34  epsilon <- (Gamma^2) / lambda
35
36  result <- list(F_principal = F,
37                 G_principal = G,
38                 F_standard = F_std,
39                 G_standard = G_std,
40                 Inercja_całkowita = lambda,
41                 Inercja_główna = epsilon)
42
43  if (plot_type %in% c("principal-principal", "standard-standard", "
    standard-principal", "principal-standard")) {
44    data_F <- data.frame(Dimension1 = F[, 1], Dimension2 = F[, 2], label
    = rownames(x)[1:(nrows-1)], group = "F")
45    data_G <- data.frame(Dimension1 = G[, 1], Dimension2 = G[, 2], label
    = colnames(x)[1:(ncols-1)], group = "G")
46    data_F_std <- data.frame(Dimension1 = F_std[, 1], Dimension2 = F_std
    [, 2], label = rownames(x)[1:(nrows-1)], group = "F_std")
47    data_G_std <- data.frame(Dimension1 = G_std[, 1], Dimension2 = G_std
    [, 2], label = colnames(x)[1:(ncols-1)], group = "G_std")
48
49    if (plot_type == "principal-principal") {
50      data <- rbind(data_F, data_G)
51      title <- "Analiza Korespondencji (principal-principal) - własna
    implementacja"
52    } else if (plot_type == "standard-standard") {
53      data <- rbind(data_F_std, data_G_std)
54      title <- "Analiza Korespondencji (standard-standard) - własna
    implementacja"
55    } else if (plot_type == "standard-principal") {
56      data <- rbind(data_F_std, data_G)
57      title <- "Analiza Korespondencji (standard-principal) - własna
    implementacja"
58    } else if (plot_type == "principal-standard") {
59      data <- rbind(data_F, data_G_std)
60      title <- "Analiza Korespondencji (principal-standard) - własna
    implementacja"
61    }
62
63    plot <- ggplot(data, aes(x = Dimension1, y = Dimension2, color =
    group, label = label, shape = group)) +
64      geom_point(size = 3) +
65      geom_text(size = 4.5, hjust = -0.4, vjust = -0.4) +
66      geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
67      geom_vline(xintercept = 0, linetype = "dashed", color = "black") +
68      labs(x = "Wymiar 1", y = "Wymiar 2", title = title) +
69      theme_minimal() +
70      theme(axis.title = element_text(size = 14),
71            axis.text = element_text(size = 12),
72            legend.text = element_text(size = 12),
73            legend.title = element_text(size = 14),
74            plot.title = element_text(size = 16, hjust = 0.5)) +
75      xlim(c(min(data$Dimension1) - 0.1, max(data$Dimension1) + 0.1)) +
76      ylim(c(min(data$Dimension2) - 0.1, max(data$Dimension2) + 0.1)) +
77      scale_color_manual(values = c("F" = "blue", "G" = "red", "F_std" =
    "blue", "G_std" = "red")) +
78      scale_shape_manual(values = c(16, 17)) +
79      guides(color = "none", shape = "none")
80
```

```
    print(plot)

    filename <- sprintf("%s.png", plot_type)
    ggsave(filename = filename, plot = plot, device = "png", height = 6,
   width = 9)
  }


  return(result)
}


plot_types <- c("principal-principal", "standard-standard", "standard-
   principal", "principal-standard")

for (plot_type in plot_types) {
  result <- correspondence_analysis(table, plot_type = plot_type)
  print(result)
```

W wyniku działania powyższej funkcji otrzymałyśmy 4 wykresy obrazujące analizę korespondencji dla grupy F i G. Na każdym z nich czerwone trójkąty reprezentują kategorie zmiennej wierszowej (zadowolenie z wynagrodzenia), niebieskie kropki przedstawiają kategorie zmiennej kolumnowej (staż pracy).

Wykresy te mają typy przedstawione w poniższej tabeli:

Typ wykresu	Typ dla wierszy	Typ dla kolumn
Standard	Standard	Standard
Row Principal	Principal	Standard
Colum Principal	Standard	Principal
Principal	Principal	Principal

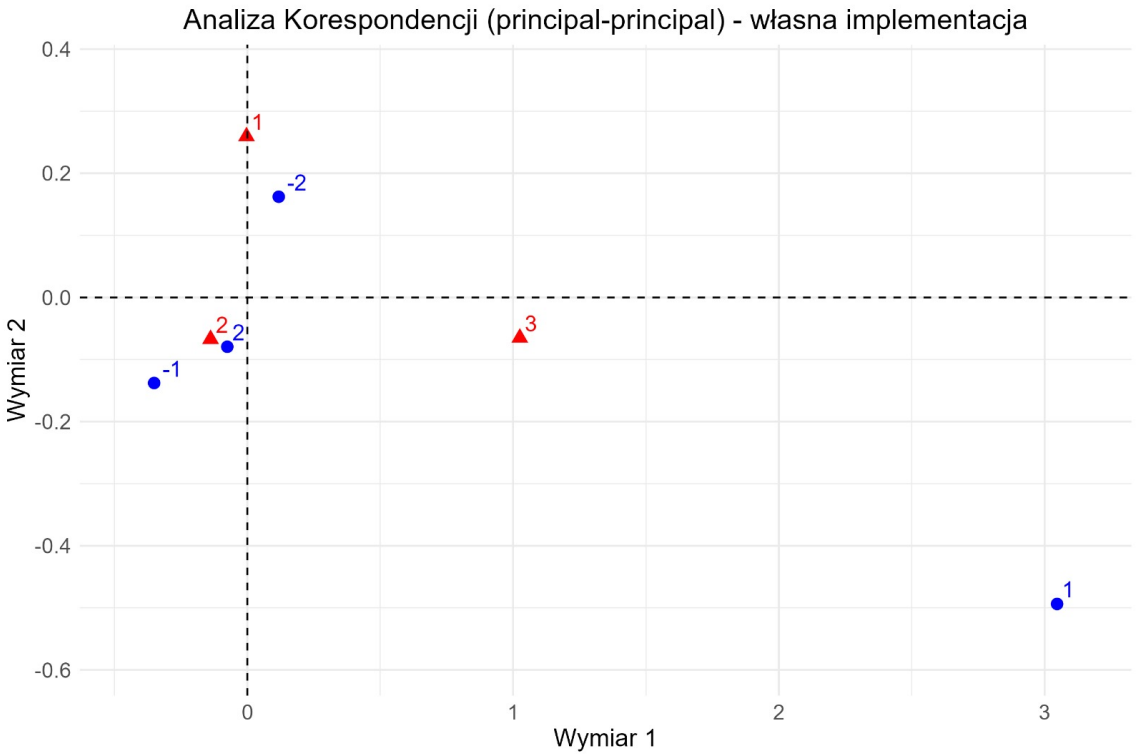
Tabela 11: Typy wykresów obrazujących analizę korespondencji

Każdy z typów wykresów ma swoje wady i zalety. Z wykresu typu Standard nie można jednoznacznie interpretować związku między kategoriami wierszy i kolumn. Odległości między kategoriami są powiększone, co uniemożliwia jasną interpretację tych odległości.

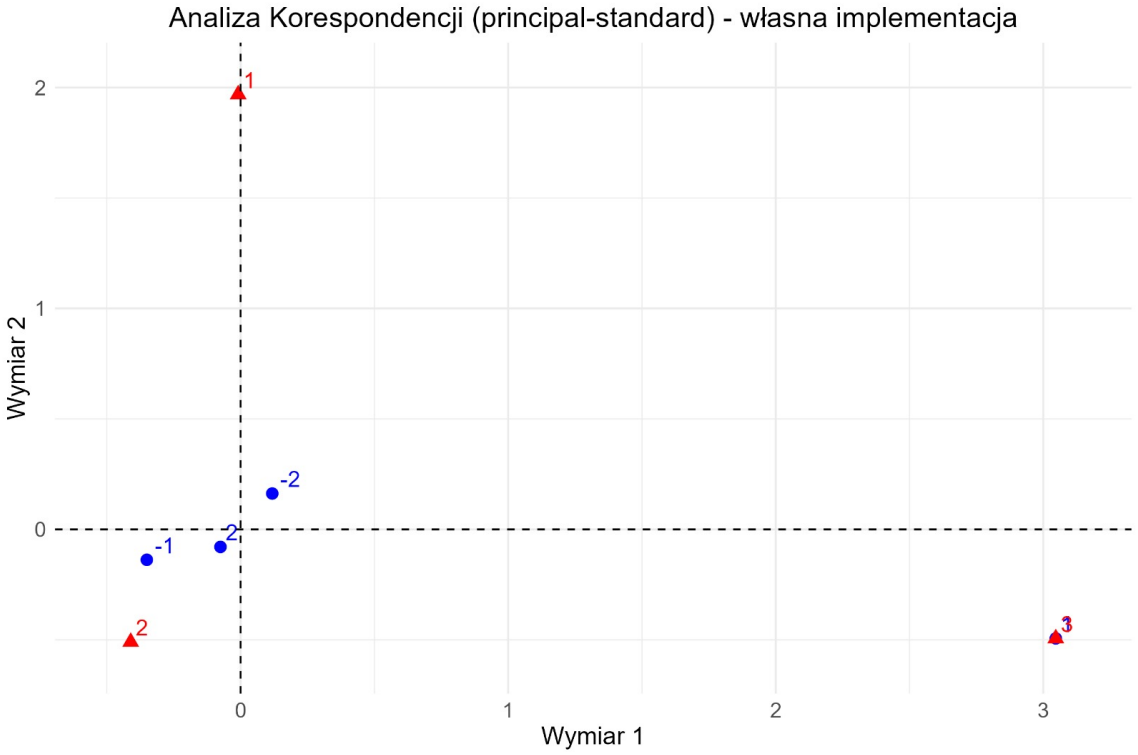
W wykresach typu Row Principal odległości między kategoriami wierszy wskazują na podobieństwo ich profili, natomiast podobna interpretacja nie jest możliwa dla kategorii kolumn, gdyż ich odległości są powiększone. Związek między kategoriami wierszy i kolumn interpretuje się za pomocą iloczynu skalarnego.

W przypadku wykresu typu Column Principal odległości między kategoriami kolumn wskazują na podobieństwo ich profili, ale taka interpretacja nie jest możliwa dla kategorii wierszy, ponieważ ich odległości są powiększone. Związek między kategoriami wierszy i kolumn również interpretuje się za pomocą iloczynu skalarnego.

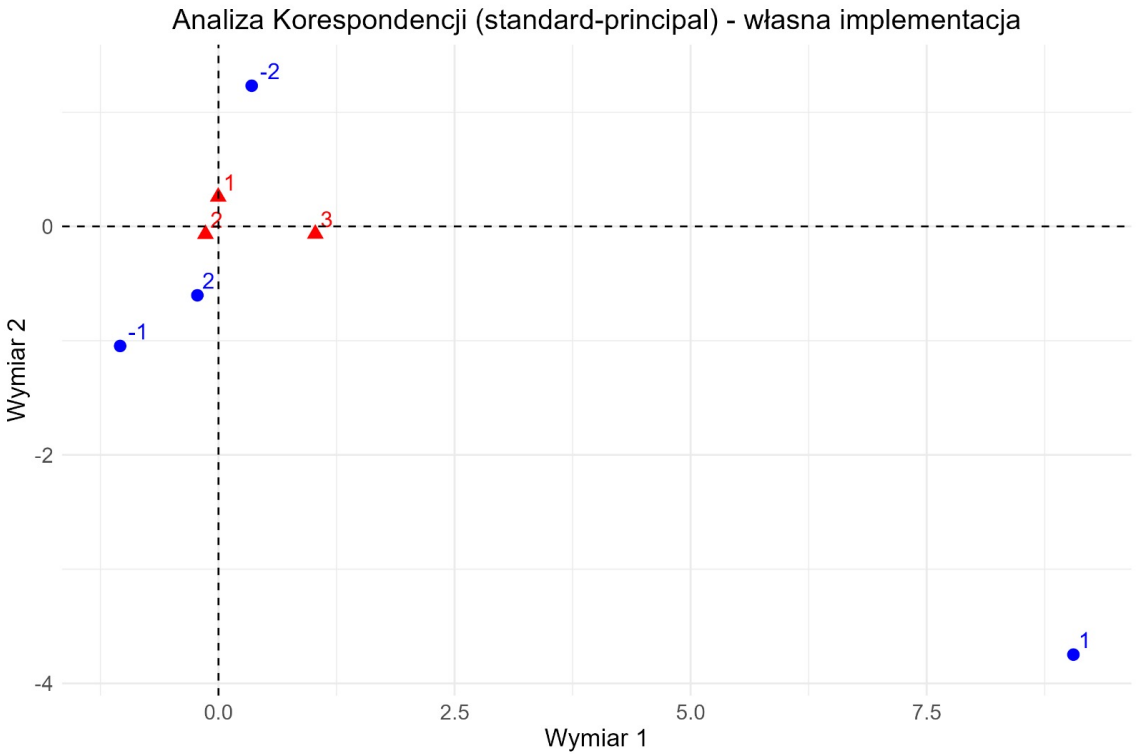
Natomiast w wykresie typu Principal odległości między kategoriami kolumn wskazują na podobieństwo ich profili, tak samo jak w przypadku kategorii wierszy. Nie można jednak jednoznacznie interpretować odległości między kategoriami wierszy i kolumn.



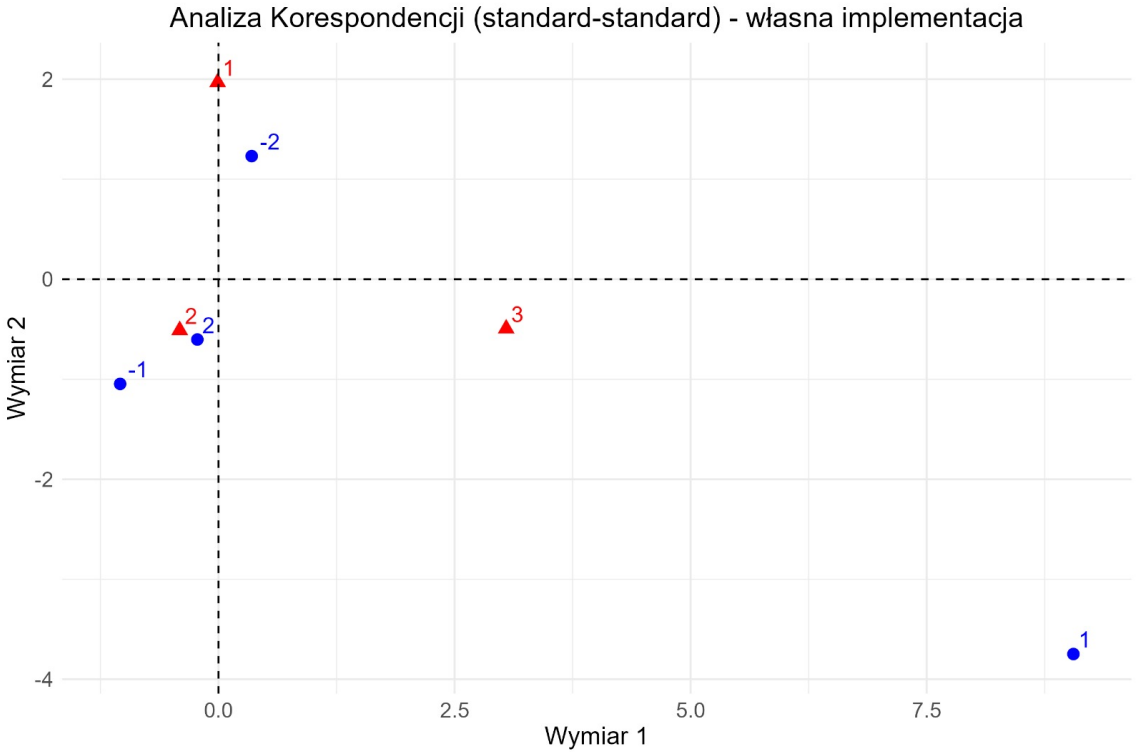
Rysunek 3: Analiza korespondecji - własna implementacja, wykres typu Principal



Rysunek 4: Analiza korespondecji - własna implementacja, wykres typu Row Principal



Rysunek 5: Analiza korespondecji - własna implementacja, wykres typu Column Principal



Rysunek 6: Analiza korespondecji - własna implementacja, wykres typu Standard

Na wykresie 3 (typu Principal) widzmy, że kategoria 1 jest wyraźnie oddalona od innych kategorii zmiennej wierszowej, co sugeruje, że jest ona odmienna od pozostałych. Pozostałe kategorie zmiennej wierszowej (-2, -1, 2) są zgrupowane bliżej siebie. Kategorie zmiennej kolumnowej są blisko siebie przez co możemy wnioskować, że kategorie stażu pracy są dość podobne w kontekście ich związku z zadowoleniem z wynagrodzenia.

Na wykresie 4 (typu Row Principal) widzimy większe zróżnicowanie rozproszenia punktów. Kategoria 3 zmiennej kolumnowej jest znacznie bardziej oddalona niż w przypadku pozostałych wykresów. Na tym wykresie także widzimy, że kategoria 1 zmiennej wierszowej jest odmienna.

Na wykresie 5 (typu Column Principal) widzimy, że wartości zmiennej wierszowej są bardziej skoncentrowane w pobliżu początku układu współrzędnych. Tutaj także kategoria 1 jest znacznie oddalona, co sugeruje jej wyjątkowość. Natomiast kategorie zmiennej kolumnowej mają większe odchylenie, co może oznaczać większe różnice w poziomach zadowolenia w zależności od stażu pracy.

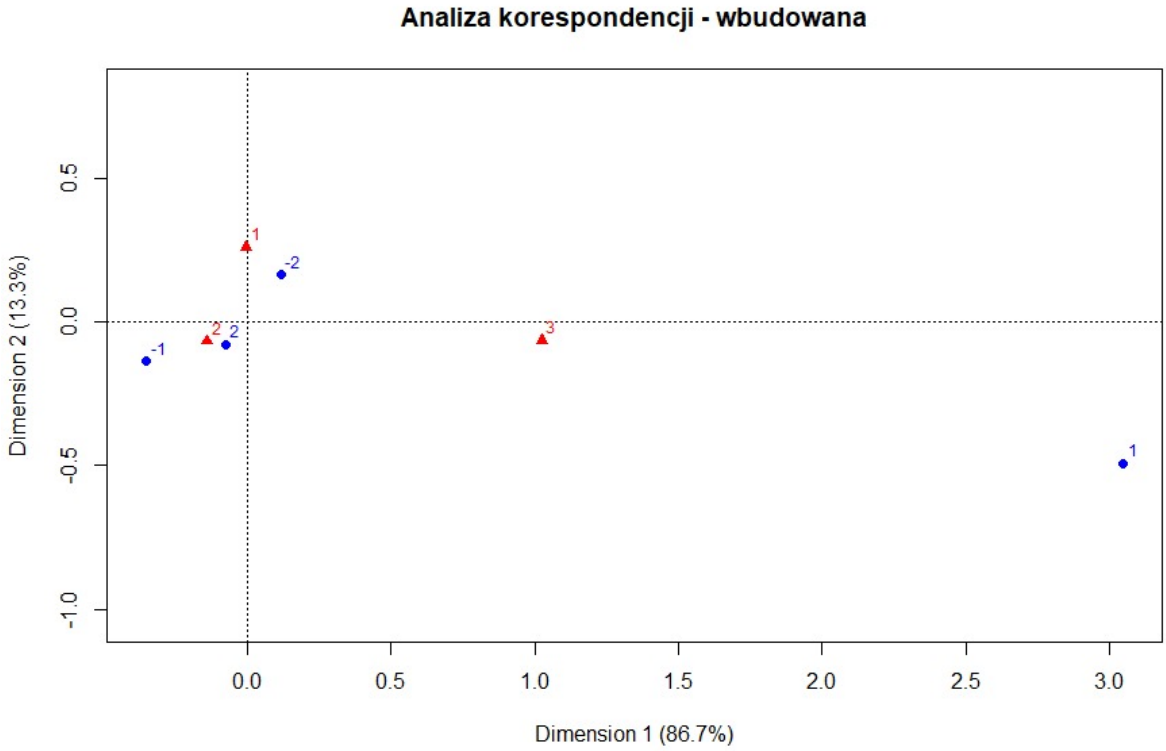
Na wykresie 6 (typu Standard) kategoria 1 zmiennej wierszowej, poodbnie jak w przypadku wykresów 3, 4, 5 znajduje się daleko na osi Y, co sugeruje jej znaczną różnicę w kontekście wpływu na ogólny rozkład danych. Widzimy także, że większość kategorii zmiennej kolumnowej jest blisko osi Y (wartości bliskie zeru), co oznacza, że mają one mniejszy wpływ na różnicowanie danych w porównaniu do kategorii 1. Standaryzacja pomaga zidentyfikować kategorie, które mają mniej ekstremalne współrzędne, sugerując ich podobieństwo lub mniejszą zmienność.

W celu sprawdzenia poprawności naszej implementacji wykorzystałyśmy bibliotekę *ca* oraz funkcje w niej wbudowane.

```
# porównanie z wbudowana
library(ca)
help(ca)

ca_built <- ca(table(data$PYT_2, data$STAŻ))
ca_built
plot(ca_built , main="Analiza korespondencji - wbudowana")
```

Z pomocą funkcji wbudowanej narysowałyśmy wykres typu Principal, który jest widoczny poniżej.



Rysunek 7: Analiza korespondecji - wbudowana funkcja

Analizując wygląd wykresów 3 oraz 7 możemy stwierdzić, że zaimplementowana przez nas funkcja najprawdopodobniej działa poprawnie, ponieważ wykres ten jest bardzo zbliżony do wykresu 3.

Policzyłyśmy także, korzystając z zaimplementowanej przez nas funkcji, inercję całkowitą i główną. Są one równe odpowiednio: inercja całkowita: 0.130597, inercja główna: 0.8449, 0.1002, $4.54931 * 10^{-33}$. Widzimy, że wartość inercji całkowitej jest dość mała co sugeruje, że punkty są mało rozproszone. Inercja główna to suma inercji całkowitych pokryta przez $i - ty$ zestaw wektorów, czym jest ona wyższa tym bardziej udana jest analiza składowych głównych.

8 Zadania dodatkowe

8.1 Zadanie *1

Napisz funkcję, która dla dwóch wektorów danych oblicza wartość poziomu krytycznego (p-value) w teście opartym na korelacji odległości. Następnie dla wygenerowanych danych zweryfikuj hipotezę o niezależności przy użyciu napisanej funkcji.

W celu weryfikacji hipotezy o niezależności dwóch wektorów danych, wygenerowałyśmy dwie próbki losowe z rozkładu normalnego o średniej 0 i odchyleniu standardowym 1 oraz o długości 100. Parametry użyte w analizie to:

- Długość wektorów: 100
- Liczba permutacji dla testu: 10000

- Poziom istotności: 0.05

Następnie przeprowadziłyśmy analizę polegającą na:

1. Obliczanie macierzy odległości (funkcja *compute_distance_matrix*) - funkcja ta oblicza macierz odległości dla danego wektora.
2. Średnie centrowanie macierzy (funkcja *center_distance_matrix*) - funkcja ta centruje macierz odległości poprzez odejmowanie średnich wierszowych i kolumnowych oraz dodawanie średniej globalnej.
3. Obliczanie korelacji odległości (funkcja *calculate_distance_correlation*) - funkcja ta oblicza współczynnik korelacji odległości dla dwóch wektorów.
4. Test korelacji odległości (funkcja *distance_correlation_test*) - funkcja ta przeprowadza test permutacyjny, obliczając wartość p dla zaobserwowanego współczynnika korelacji odległości.

Kod jest widoczny poniżej.

```r
# Obliczanie macierzy odległości
compute_distance_matrix <- function(v) {
  n <- length(v)
  dist_matrix <- matrix(0, n, n)
  for (i in 1:n) {
    for (j in 1:n) {
      dist_matrix[i, j] <- abs(v[i] - v[j])
    }
  }
  return(dist_matrix)
}

# Średnie centrowanie macierzy
center_distance_matrix <- function(dm) {
  n <- nrow(dm)
  row_means <- colMeans(dm)
  col_means <- rowMeans(dm)
  grand_mean <- mean(dm)

  for (i in 1:n) {
    for (j in 1:n) {
      dm[i, j] <- dm[i, j] - row_means[i] - col_means[j] + grand_mean
    }
  }
  return(dm)
}

# Obliczanie korelacji odległości
calculate_distance_correlation <- function(X, Y) {
  n <- length(X)
  A <- center_distance_matrix(compute_distance_matrix(X))
  B <- center_distance_matrix(compute_distance_matrix(Y))

  A_B <- sum(A * B) / n^2
  A_A <- sum(A * A) / n^2
  B_B <- sum(B * B) / n^2

  result <- sqrt(A_B / sqrt(A_A * B_B))
```

```r
  return(result)
}

# Funkcja testująca
distance_correlation_test <- function(x, y, B = 10000) {
  observed_dcor <- calculate_distance_correlation(x, y)
  n <- length(x)
  permuted_dcor <- numeric(B)

  for (i in 1:B) {
    permuted_dcor[i] <- calculate_distance_correlation(x, sample(y))
  }

  p_value <- mean(abs(permuted_dcor) >= abs(observed_dcor))
  return(list(statistic = observed_dcor, p_value = p_value))
}

# Przykładowe dane
set.seed(123)
alpha <- 0.05
x <- rnorm(100)
y <- rnorm(100)

# Test
result <- distance_correlation_test(x, y)

# Wyniki testu z biblioteki energy
dcor_built <- dcor(x, y)
dcor_result <- dcor.test(x, y, R = 10000)

# Tworzenie tabeli z wynikami
results_table <- data.frame(
  Metoda = c("Własna", "Wbudowana", "Różnica"),
  d_korelacja = c(result$statistic, dcor_built, abs(dcor_built - result$
    statistic)),
  `p-wartość` = c(result$p_value, dcor_result$p.value, abs(dcor_result$p.
    value - result$p_value)),
  Odrzucenie_H0 = c(result$p_value < alpha, dcor_result$p.value < alpha,
    NA)
)

print(results_table)
```

Powyżej została przedstawiona zaimplementowana przez nas funkcja oraz jej porównanie z wbudowanymi funkcjami *dcor* oraz *dcor.test* z biblioteki *energy*.

W wyniku jej działania otrzymaliśmy tabelę z wynikami, która jest widoczna poniżej:

Metoda	d-korelacja	p-wartość	Odrzucenie H0
Własna	0.1565	0.663600	FALSE
Wbudowana	0.1565	0.659034	FALSE
Różnica	$1.9429 * 10^{-16}$	0.004566	

Tabela 12: Porównanie własnej implementacji z funkcją wbudowaną

Jak możemy zauważyć w tabeli 12 wartości d-korelacji obliczone za pomocą własnej implementacji oraz wbudowanej funkcji są identyczne (równe około 0.1565). Wskazuje to na poprawność implementacji algorytmu obliczania korelacji odległości we własnej funkcji.

P-wartości dla obu metod są do siebie dość zbliżone: 0.663600 dla własnej funkcji i około 0.659034 dla funkcji wbudowanej. Ich różnica jest bardzo mała (0.004566), co jest akceptowalne przy losowym próbkowaniu danych. Na podstawie otrzymanych p-wartości, które w obu przypadkach są większe od poziomu istotności równego 0.05 możemy wnioskować, że nie ma podstaw do odrzucenia hipotezy zerowej.

Podsumowując, możemy wnioskować, że napisana przez nas funkcja działa prawidłowo i dostarcza wyniki porównywalne z wbudowaną. Różnice w p-wartościach są niewielkie i mogą wynikać z losowego charakteru permutacji. Możemy więc stwierdzić, że własna funkcja jest poprawnie zaimplementowana i może być używana do testowania niezależności przy użyciu korelacji odległości. Główny wniosek, czyli brak podstaw do odrzucenia hipotezy zerowej o niezależności, jest spójny w obu metodach.

8.2 Zadanie *2

Dla zadanych π_1 oraz π_2 pokaż, że wartość ryzyka względnego (RR) nie jest bardziej oddalona od wartości 1 (wartość odpowiadająca niezależności) niż wartość odpowiadającego ilorazu szans (OR).

Dane są:

- π_1 - prawdopodobieństwo zdarzenia w grupie eksponowanej,
- π_2 - prawdopodobieństwo zdarzenia w grupie nieeksponowanej.

Do rozwiązania zadania wykorzystałyśmy wzory na ryzyko względne (RR) i iloraz szans (OR), które były podane na wykładzie i laboratoriach:

$$\mathrm{RR} = \frac{\pi_1}{\pi_2}$$

$$\mathrm{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Chcemy pokazać, że:

$$|\mathrm{RR} - 1| \leq |\mathrm{OR} - 1|$$

Rozpoczęłyśmy od przekształcenia wzoru na ryzko względne (RR):

$$|\mathrm{RR} - 1| = \left|\frac{\pi_1}{\pi_2} - 1\right|,$$

oraz na iloraz szans (OR):

$$|\mathrm{OR} - 1| = \left|\frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} - 1\right|$$

Następnie uprościłyśmy wyrażenie wewnątrz wartości bezwzględnej dla OR:

$$\mathrm{OR} - 1 = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} - 1 = \frac{\pi_1(1-\pi_2) - \pi_2(1-\pi_1)}{\pi_2(1-\pi_1)} = \frac{\pi_1 - \pi_1\pi_2 - \pi_2 + \pi_1\pi_2}{\pi_2(1-\pi_1)} = \frac{\pi_1 - \pi_2}{\pi_2(1-\pi_1)}$$

W związku z tym wartość bezwzględna dla OR staje się:

$$|\mathrm{OR} - 1| = \left|\frac{\pi_1 - \pi_2}{\pi_2(1-\pi_1)}\right|$$

Następnie porównałyśmy ze sobą oba wyrażenia:

$$\left|\frac{\pi_1}{\pi_2} - 1\right| \quad \text{i} \quad \left|\frac{\pi_1 - \pi_2}{\pi_2(1-\pi_1)}\right|$$

Chcemy pokazać, że $\left|\frac{\pi_1}{\pi_2} - 1\right| \leq \left|\frac{\pi_1-\pi_2}{\pi_2(1-\pi_1)}\right|$.
Przekształcając $\frac{\pi_1}{\pi_2} - 1$:

$$\left|\frac{\pi_1}{\pi_2} - 1\right| = \left|\frac{\pi_1 - \pi_2}{\pi_2}\right|$$

W ten sposób uzyskaliśmy do udowodnienia poniższą nierówność:

$$\left|\frac{\pi_1 - \pi_2}{\pi_2}\right| \leq \left|\frac{\pi_1 - \pi_2}{\pi_2(1-\pi_1)}\right|$$

Wiemy, że π_1 i π_2 są prawdopodobieństwami, $0 \leq \pi_1, \pi_2 \leq 1$. W związku z tym wyrażenie $(1-\pi_1)$ jest zawsze dodatnie i mniejsze lub równe 1. Dlatego dzielenie przez $(1-\pi_1)$ daje:

$$\left|\frac{\pi_1 - \pi_2}{\pi_2}\right| \leq \left|\frac{\pi_1 - \pi_2}{\pi_2} \cdot \frac{1}{1-\pi_1}\right|$$

Wiemy także, że $\frac{1}{1-\pi_1} \geq 1$, w związku z tym potwierdziłyśmy, że:

$$\left|\frac{\pi_1 - \pi_2}{\pi_2}\right| \leq \left|\frac{\pi_1 - \pi_2}{\pi_2(1-\pi_1)}\right|$$

Stąd:

$$|\text{RR} - 1| \leq |\text{OR} - 1|$$

Pokazałyśmy, że ryzyko względne (RR) nie jest bardziej oddalone od wartości 1 niż iloraz szans (OR). To kończy dowód.

8.3 Zadanie *3

Niech D oznacza posiadanie pewnej choroby, a E pozostawanie wystawionym na pewny czynnik ryzyka. W badaniach epidemiologicznych definiuje się miarę AR nazywaną ryzykiem przypisanym *(ang. attributable risk)*.

a) Niech $P(E') = 1 - P(E)$, wówczas $AR = [P(D) - P(D|E')]/P(D)$. Wyjaśnij interpretację miary na podstawie wzoru.

b) Pokaż, że AR ma związek z ryzykiem względnym, tzn.:

$$AR = [P(E)(RR-1)]/[1 + P(E)(RR-1)].$$

W celu wyjaśnienia interpretacji miary z podpunktu a), w pierwszej kolejności wyjaśnimy oznaczenia użyte we wzorze $P(E') = 1 - P(E)$:

- $P(D)$ - prawdopodobieństwo posiadania choroby,

- $P(D|E')$ - prawdopodobieństwo posiadania choroby przy braku wystawienia na czynnik ryzyka E.

Korzystając ze wzoru z treści zadania $AR = [P(D) - P(D|E')]/P(D)$ możemy interpretować, że miara AR reprezentuje proporcję całkowitego ryzyka choroby ($P(D)$), która jest przypisana ekspozycji na czynnik ryzyka (E).

Następnie przeszłyśmy do podpunktu b). Na początek przypomnimy, że ryzyko względne (RR) definiujemy jako:

$$RR = \frac{P(D|E)}{P(D|E')}$$

Widzimy, że:

$$P(D) = P(D|E)P(E) + P(D|E')P(E')$$

gdzie:

- $P(E') = 1 - P(E)$
- $P(D|E)$ - prawdopodobieństwo choroby przy ekspozycji.
- $P(D|E')$ - prawdopodobieństwo choroby przy braku ekspozycji.

W tej części zadania chciałyśmy pokazać, że:

$$AR = \frac{P(E)(RR-1)}{1 + P(E)(RR-1)}$$

Zaczęłyśmy od wzoru na AR (z laboratoriów):

$$AR = \frac{P(D) - P(D|E')}{P(D)}$$

Następnie podstawiłyśmy $P(D)$ do licznika:

$$AR = \frac{P(D|E)P(E) + P(D|E')P(E') - P(D|E')}{P(D)}$$

Po jego zredukowaniu otrzymałyśmy:

$$AR = \frac{P(D|E)P(E) - P(D|E')P(E) + P(D|E') - P(D|E')}{P(D)} = \frac{P(E)(P(D|E) - P(D|E'))}{P(D)}$$

Dodatkowo, wiemy że $RR = \frac{P(D|E)}{P(D|E')}$, więc:

$$P(D|E) = RR \cdot P(D|E')$$

W kolejnym kroku podstawiłyśmy $P(D|E)$:

$$AR = \frac{P(E)(RR \cdot P(D|E') - P(D|E'))}{P(D)} = \frac{P(E)P(D|E')(RR-1)}{P(D)}$$

Oraz podstawiłyśmy $P(D)$ i RR do mianownika:

$$P(D) = P(D|E)P(E) + P(D|E')P(E') = RR \cdot P(D|E') \cdot P(E) + P(D|E') \cdot (1 - P(E)) =$$

$$= P(D|E') \cdot [RR \cdot P(E) + (1 - P(E))] = P(D|E') \cdot [1 + P(E) \cdot (RR-1)]$$

Następnie podstawiłyśmy wartość $P(D)$ obliczoną powyżej do mianownika wzoru na AR:

$$AR = \frac{P(E)P(D|E')(RR-1)}{P(D|E') \cdot [1 + P(E)(RR-1)]} = \frac{P(E)(RR-1)}{1 + P(E)(RR-1)}$$

W ten sposób pokazałyśmy, że ryzyko przypisane (AR) można wyrazić w zależności od ryzyka względnego (RR):

$$AR = \frac{P(E)(RR-1)}{1 + P(E)(RR-1)}$$

9 Źródła

- https://statsandr.com/blog/fisher-s-exact-test-in-r-independence-test-for-a-small-sample/
- https://www.scribbr.com/statistics/chi-square-tests/
- https://www.sciencedirect.com/science/article/pii/S0167715218304048
- https://projecteuclid.org/journals/annals-of-statistics/volume-35/issue-6/Measuring-and-testing-dependence-by-correlation-of-distances/10.1214/009053607000000505.full
- https://link.springer.com/book/10.1007/978-1-4757-2346-5
- https://www.r-project.org/other-docs.html
- https://www.rdocumentation.org/