Analiza danych ankietowych - sprawozdanie 3

"Lista zadań nr 3"

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 **3**

2 Użyte biblioteki **3**

3 Wczytywanie danych **3**

4 Część I oraz II **3**
4.1 Zadanie 1 3
4.2 Zadanie 2 5
4.3 Zadanie 3 6
4.4 Zadanie 4 10
4.5 Zadanie 5 11

5 Część III **14**
5.1 Zadanie 6 14
5.2 Zadanie 7 17

6 Część IV oraz V **18**
6.1 Zadanie 8 18
6.2 Zadanie 9 23

7 Źródła **26**

1 Wstęp

W poniższym sprawozdaniu zostały przedstawione wyniki listy zadań nr 3 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(exact2x2)
library(gnm)
```

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")

data$CZY_ZADOW_2 <- ifelse(data$PYT_3 %in% c("-2", "-1"), "NIE", "TAK")
```

4 Część I oraz II

4.1 Zadanie 1

Napisz funkcję, która zwraca p-wartość w omówionym na wykładzie warunkowym teście symetrii dla tabel 2×2.

W tym zadaniu napisałyśmy funkcję *symetric_test*, która służy do obliczenia p-wartości w warunkowym teście symetrii dla tabeli $2x2$. Test symetrii sprawdza, czy rozkład wartości w tabeli jest symetryczny. Kod z zaimplementowaną funkcją widoczny jest poniżej.

```
symetric_test <- function(table) {
  y_1_2 <- table[1, 2]
  y_2_1 <- table[2, 1]

```

```r
  n_star <- y_1_2 + y_2_1

  y_L <- min(y_1_2, n_star - y_1_2)
  y_P <- max(y_1_2, n_star - y_1_2)

  if (y_L == y_P) {
    return(1)
  }

  else {
  sum1 <- sum(sapply(0:y_L, function(i) dbinom(i, n_star, 1/2)))
  sum2 <- sum(sapply(y_P:n_star, function(i) dbinom(i, n_star, 1/2)))

  p_value <- sum1 + sum2

  return(p_value)
  }
}
```

Następnie przeszłyśmy do sprawdzenia poprawności powyższego kodu. W tym celu zbadałyśmy symetrię 2 przykładowych tabel.

10	20
5	15

Tabela 1: Przykładowa tabela 1

10	5
5	10

Tabela 2: Przykładowa tabela 2

```r
# Przykładowe tabele, w których sprawdzamy czy występuje w nich brzegowa
    jednorodność
# Przykładowa tabela 1
table_1 <- matrix(c(10, 20, 5, 15), nrow = 2, byrow = TRUE)
p_value_1 <- symetric_test(table_1)
print(p_value_1)

# Przykładowa tabela 2
table_2 <- matrix(c(10, 5, 5, 10), nrow = 2, byrow = TRUE)
p_value_2 <- symetric_test(table_2)
print(p_value_2)
```

W wyniku działania powyższego kodu otrzymałyśmy następujące p-wartości:

- Dla tabeli 1: p-wartość ≈ 0.004077,
- Dla tabeli 2: p-wartość $= 1$.

Analizując powyższe p-wartości możemy wnioskować, że na poziomie istotności $\alpha = 0.05$, w przypadku tabeli 1 mamy podstawy do odrzucenia hipotezy zerowej (ponieważ p-wartość < 0.05) o symetrii, która jest równoważna hipotezie o brzegowej jednorodności. Natomiast w przypadku tabeli 2 nie mamy podstaw do jej odrzucenia, co pozwala nam wnioskować, że rozkład wartości w tabeli 2 jest symetryczny. Na tej podstawie możemy stwierdzić, że nasza funkcja najprawdopodobniej działa poprawnie. W związku z tym będziemy mogli wykorzystać ją w zadaniu 2.

4.2 Zadanie 2

W tabeli 3 umieszczono dane dotyczące reakcji na lek po godzinie od jego przyjęcia dla dwóch różnych leków przeciwbólowych stosowanych w migrenie. Leki zostały zaaplikowane grupie pacjentów w dwóch różnych atakach bólowych. Na podstawie danych zweryfikuj hipotezę, że leki te są jednakowo skuteczne korzystając z testu

- McNemara z poprawka na ciągłość,
- warunkowego (korzystając z funkcji zadeklarowanej w zadaniu 1.).

Reakcja na lek A	Reakcja na lek B	
	Negatywna	Pozytywna
Negatywna	1	5
Pozytywna	2	4

Tabela 3: Dane do zadania 2.

W celu rozwiązania tego zadania użyłyśmy funkcji *mcnemar.exact* z biblioteki *exact2x2*. Dodanie poprawki na ciągłość sprawia, że test jest bardziej konserwatywny, co oznacza, że jest mniej skłonny do odrzucenia hipotezy zerowej.

Natomiast test warunkowy, oparty na rozkładzie dwumianowym, ocenia, czy rozkład wartości w tabeli 2x2 jest symetryczny, co odpowiada sprawdzeniu, czy skuteczność obu leków jest taka sama.

W celu obliczenia p-wartości dla obu tych testów napisałyśmy kod widoczny poniżej. Jako poziom istotności przyjęłyśmy $\alpha = 0.05$.

```r
# tabela z zadania
medicine_data <- matrix(c(1, 5, 2, 4), nrow = 2, byrow = TRUE)

rownames(medicine_data) <- c("Negatywna", "Pozytywna") # lek A
colnames(medicine_data) <- c("Negatywna", "Pozytywna") # lek B

reaction_table <- as.table(medicine_data)
print(reaction_table)
```

```r
# test McNemara z poprawka na ciągłość,
mcNemar <- mcnemar.exact(medicine_data)
cat("p-wartość dla testu McNemara:", mcNemar$p.value , "\n")
```

```r
# test warunkowy (funkcja zadeklarowanej w zadaniu 1).
symetric_p_val <- symetric_test(medicine_data)
cat("p-wartość dla testu symetrii 2x2:", symetric_p_val  , "\n")
```

```r
result <- data.frame(
  "test" = c("McNemar z poprawka", "test warunkowy"),
  "p-wartość" = c(mcNemar$p.value, symetric_p_val)
)
result
```

W wyniku przeprowadzenia dwóch testów statystycznych na danych dotyczących skuteczności dwóch leków przeciwbólowych, otrzymano następujące p-wartości:

test	p.wartość
McNemar z poprawka	0.453125
test warunkowy	0.453125

Tabela 4: Obliczone p-wartości

W tabeli 4 możemy zauważyć, że p-wartości dla obu testów są znacznie wyższe niż ustalony poziom istotności równy 0.05. Oznacza to, że zarówno test McNemara z poprawką na ciągłość, jak i warunkowy test symetrii wskazują, że nie mamy podstaw, aby odrzucić hipotezę zerową mówiącą o tym, że oba leki są jednakowo skuteczne. Innymi słowy, na podstawie dostępnych danych nie możemy stwierdzić, że istnieje istotna różnica w skuteczności między dwoma badanymi lekami.

4.3 Zadanie 3

Przeprowadź symulacje w celu porównania mocy testu Z i testu Z_0 przedstawionych na wykładzie. Rozważ różne długości prób.

Przyjęłyśmy długości prób $n \in \{20, 50, 100, 1000\}$. Poniżej widoczny jest kod służący do porównania mocy testów Z i Z_0.

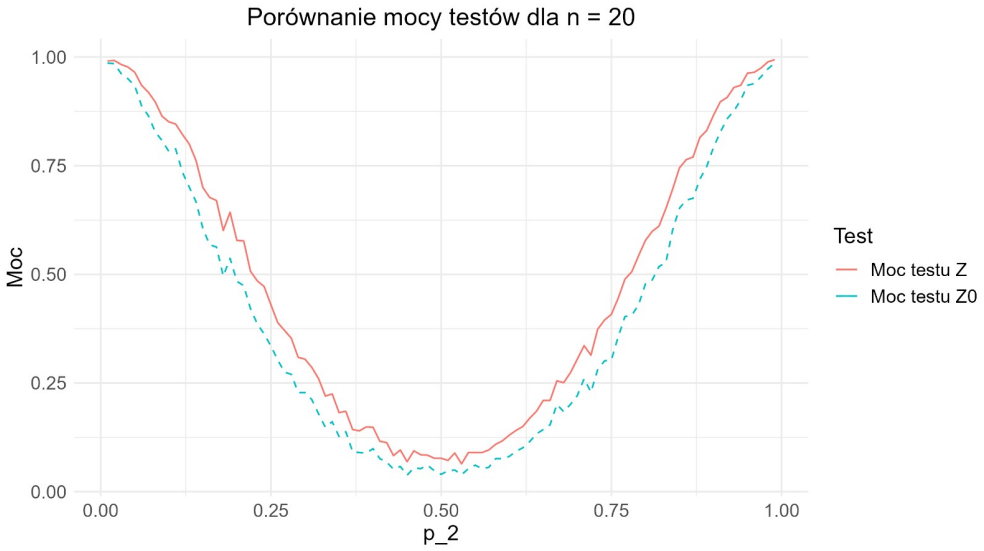
```r
# Funkcja tworząca tablicę
table_making <- function(p_1, p_2, n) {
  X <- rbinom(n, 1, p_1)
  Y <- rbinom(n, 1, p_2)

  y <- table( factor(X, levels = c("0", "1")),
              factor(Y, levels = c("0", "1")))

  y <- rbind(y, colSums(y))
  y <- cbind(y, rowSums(y))

  return(y)
}

# Funkcja przeprowadzająca testy Z i Z0
Z_Z_0_tests <- function(y) {
  nrows <- dim(y)[1]
  ncols <- dim(y)[2]

  n <- y[nrows, ncols]

  p <- y/n

  D <- (y[1, 2] - y[2, 1])/n

  # test Z
  p_1_plus <- p[1, ncols]
  p_plus_1 <- p[nrows, 1]
  sigma2 <- (p_1_plus*(1 - p_1_plus) +
             p_plus_1*(1 - p_plus_1) -
             2*(p[1, 1]*p[2, 2] - p[1, 2]*p[2, 1]) ) / n

  Z <- D/sqrt(sigma2)
  p_val <- 2*(1 - pnorm(abs(Z)))

  # test Z_0
```

```r
  Z_0 <- (y[1, 2] - y[2, 1])/sqrt(y[1, 2] + y[2, 1])
  p_val_0 <- 2*(1 - pnorm(abs(Z_0)))

  return(list(Z = Z, p_val_Z = p_val, Z_0 = Z_0, p_val_Z_0 = p_val_0))
}

# Funkcja porównująca moc testów
compare_power <- function(p_1, p_2, n, alpha = 0.05, N = 1000) {
  power_Z <- numeric(length(p_2))
  power_Z_0 <- numeric(length(p_2))

  for (i in 1:length(p_2)) {
    count_Z <- 0
    count_Z_0 <- 0

    for (k in 1:N) {
      table <- table_making(p_1, p_2[i], n)
      Z_Z_0 <- Z_Z_0_tests(table)
      p_val <- Z_Z_0$p_val_Z
      p_val_0 <- Z_Z_0$p_val_Z_0

      if (p_val < alpha) {
        count_Z <- count_Z + 1
      }
      if (p_val_0 < alpha) {
        count_Z_0 <- count_Z_0 + 1
      }
    }
    power_Z[i] <- count_Z / N
    power_Z_0[i] <- count_Z_0 / N
  }
  return(data.frame(p_2 = p_2, power_Z = power_Z, power_Z_0 = power_Z_0))
}
```

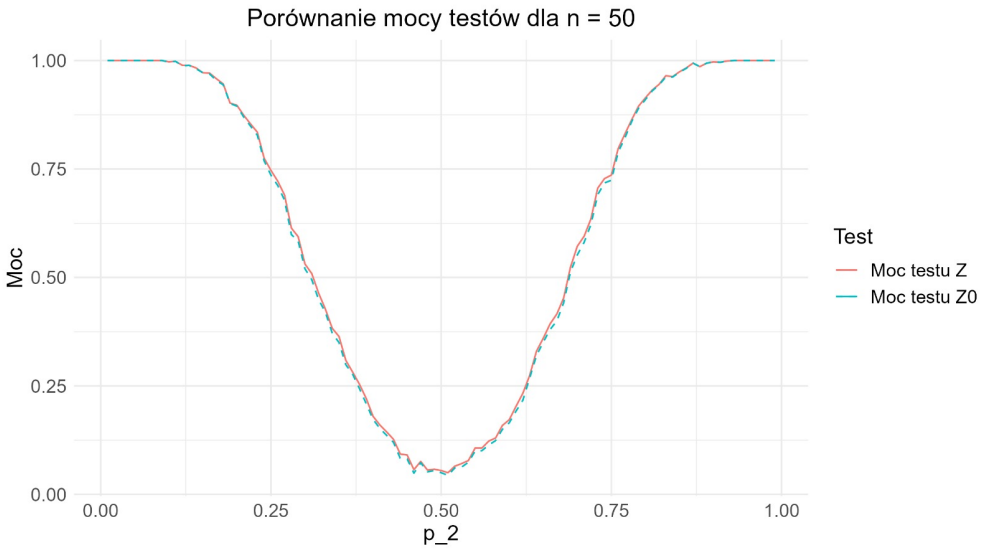
```r
# Parametry
p_1 <- 0.5
p_2 <- seq(0.01, 0.99,  by=0.01)
n <- c(20, 50, 100, 1000)

# Rysowanie wykresów
for (i in 1:length(n)) {
  results <- compare_power(p_1, p_2, n[i])
  results$n <- n[i]

  p <- ggplot(results, aes(x = p_2)) +
    geom_line(aes(y = power_Z, color = "Moc testu Z")) +
    geom_line(aes(y = power_Z_0, color = "Moc testu Z0"), linetype = "
   dashed") +
    labs(title = paste("Porównanie mocy testów dla n =", n[i]),
         x = "p_2",
         y = "Moc",
         color = "Test") +
    theme_minimal() +
    theme(axis.title = element_text(size = 14),
          axis.text = element_text(size = 12),
          legend.text = element_text(size = 12),
          legend.title = element_text(size = 14),
          plot.title = element_text(size = 16, hjust = 0.5))

```

```
  ggsave(paste("porownanie_mocy_testow_n", n[i], ".png", sep = ""), plot
    = p, width = 9, height = 5)
}
```

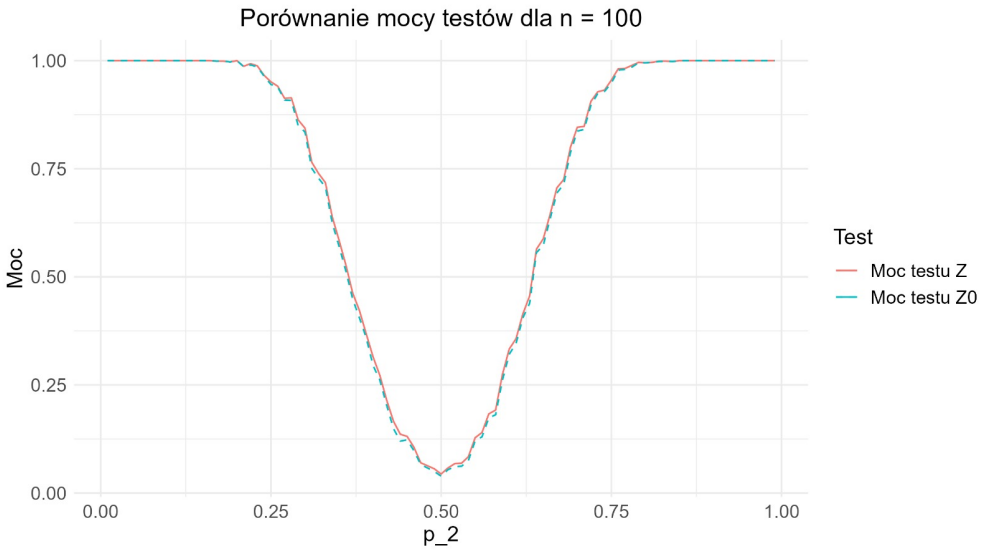
W wyniku działania powyższego kodu otrzymałyśmy 4 wykresy, które są widoczne poniżej.



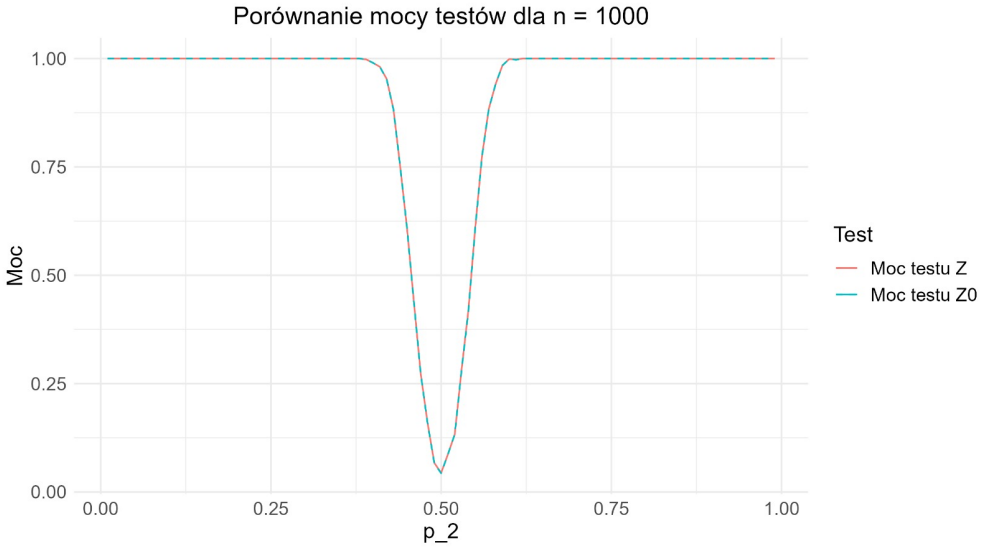
Rysunek 1: Porównanie mocy testów dla n = 20



Rysunek 2: Porównanie mocy testów dla n = 50



Rysunek 3: Porównanie mocy testów dla n = 100



Rysunek 4: Porównanie mocy testów dla n = 1000

Na wykresach 1, 2, 3, 4 możemy zauważyć, że w okolicy $p_2 = 0.5$ istnieje "dołek".

Widzimy także, że czym większe n, tym moce testów są do siebie bardziej zbliżone, różnice między nimi się niwelują, co jest zgodne z intuicją. Patrząc na wykres 1 widzimy, że test Z wydaje się być bardziej stabilny i skuteczniejszy przy małych próbach, ponieważ zachowuje on wyższą moc niż test Z_0.

4.4 Zadanie 4

Dla danych dołączonych do pierwszej listy zadań, na podstawie zmiennych **CZY_ZADW** oraz **CZY_ZADW_2**, zweryfikuj hipotezę, że zadowolenie z wynagrodzenia w pierwszym badanym okresie i po roku od pierwszego badania odpowiada modelowi symetrii. Czy na podstawie uzyskanych wyników możemy wnioskować, że poziom zadowolenia z wynagrodzenia nie uległ zmianie? Przyjmij poziom istotności 0.05.

```
table_content <- table(data$CZY_ZADOW, data$CZY_ZADOW_2)
```

	NIE	TAK
NIE	74	20
TAK	8	98

Tabela 5: Dane do zadania 4

W tabeli 5 widoczne są dane wykorzystane w poniższym zadaniu. Wykorzystałyśmy w nim funkcję *mcnemar.test* z biblioteki *stats*.

```
# testy
mcNemar_test_with <- mcnemar.test(table_content, correct = TRUE)
mcNemar_test_without <- mcnemar.test(table_content, correct = FALSE)

sym_test <- symetric_test(table_content)

table_content <- rbind(table_content, colSums(table_content))
table_content <- cbind(table_content, rowSums(table_content))
Z_Z_0_test <- Z_Z_0_tests(table_content)
```

```
# wyniki
result <- data.frame(
  "test" = c("McNemar z poprawka", "McNemar bez poprawki", "Warunkowy", "
   Z", "Z_0"),
  "p-wartość" = c(mcNemar_test_with$p.value, mcNemar_test_without$p.value
   , sym_test, Z_Z_0_test$p_val_Z, Z_Z_0_test$p_val_Z_0)
)
result
```

test	p-wartość
McNemar z poprawka	0.037635
McNemar bez poprawki	0.023342
Warunkowy	0.035698
Z	0.021589
Z_0	0.023342

Tabela 6: P-wartości dla poszczególnych testów

W tabeli 6 zostały przedstawione obliczone p-wartości dla testu McNemara z poprawką i bez niej, testu warunkowego, testu Z i testu Z_0.

P-wartość obliczona w teście McNemara z poprawką na ciągłość wynosi 0.037635 i jest mniejsza niż poziom istotności $\alpha = 0.05$, co oznacza, że mamy podstawy do odrzucenia hipotezy zerowej o braku różnicy w zadowoleniu z wynagrodzenia w dwóch okresach.

Test McNemara bez poprawki na ciągłość również wskazuje na istotną różnicę w proporcjach odpowiedzi. P-wartość 0.023342 jest mniejsza niż $\alpha = 0.05$, co również prowadzi do odrzucenia hipotezy zerowej.

Test warunkowy, podobnie jak testy McNemara, wskazuje na istotną różnicę w zadowoleniu z wynagrodzenia. P-wartość 0.035698 jest mniejsza niż 0.05, co także prowadzi do odrzucenia hipotezy zerowej.

Podobne wnioski możemy wyciągnąć z analizy p-wartości dla testów Z i Z_0, równych odpowiednio około 0.021589 i około 0.023342. W obu testach p-wartość jest mniejsza od poziomu istotności co prowadzi do odrzucenia hipotezy zerowej. W przypadku testu Z_0 otrzymaliśmy taką samą p-wartość jak w teście McNemara bez poprawki, co wynika z faktu, że p-wartość obliczona w teście Z_0 jest testem McNemara.

Możemy zauważyć, że wszystkie przeprowadzone testy (McNemara z poprawką na ciągłość, McNemara bez poprawki, warunkowy, Z, Z_0) wskazują na istotną różnicę w zadowoleniu z wynagrodzenia między dwoma badanymi okresami. Otrzymane p-wartości we wszystkich przypadkach są mniejsze niż przyjęty poziom istotności $\alpha = 0.05$, co oznacza, że mamy podstawy do odrzucenia hipotezy zerowej, która mówi o braku różnicy w poziomie zadowolenia z wynagrodzenia.

Na podstawie uzyskanych wyników możemy wnioskować, że poziom zadowolenia z wynagrodzenia uległ zmianie pomiędzy pierwszym badanym okresem, a okresem po roku. Wszystkie testy jednoznacznie wskazują na istotną statystycznie różnicę w zadowoleniu z wynagrodzenia.

4.5 Zadanie 5

W korporacji, o której mowa w zadaniu 1 z listy 1, wdrożono pewne działania w celu poprawy komfortu pracy. Następnie badaną grupę respondentów ponownie poproszono o odpowiedź na pytanie dotyczące oceny podejścia firmy do utrzymania równowagi między życiem zawodowym a prywatnym. W Tabeli 7 przedstawiono tablicę dwudzielczą uwzględniającą odpowiedzi na pytanie w obu tych okresach. Na podstawie danych zweryfikuj hipotezę, że odpowiedzi w pierwszym badanym okresie i w drugim okresie odpowiadają modelowi symetrii. Na podstawie wyników uzyskanych przy weryfikacji hipotezy dotyczącej symetrii, sformułuj wniosek dotyczący hipotezy, że ocena podejścia firmy nie uległa zmianie.

Pytanie 1	Pytanie 2				
	-2	-1	0	1	2
-2	10	2	1	1	0
-1	0	15	1	1	0
0	1	1	32	6	0
1	0	0	1	96	3
2	1	1	0	1	26

Tabela 7: Dane do zadania 5.

Do rozwiązania zadania wykorzystałyśmy funkcję *mcnemar.test()*, która służy do przeprowadzenia testu McNemara na zgodność dwóch sparowanych próbek. Skorzystałyśmy także z faktu, że test Bowkera to rozszerzenie testu McNemara stosowane do danych wielowymiarowych, gdzie wymiar jest większy od 2. W funkcji tej parametr *correct* odnosi się do zastosowania poprawki ciągłości (poprawki Yatesa) w teście McNemara. Jeśli *correct* = *TRUE* to zastosowano poprawkę na ciągłość, jeśli *FALSE* to nie zastosowano jej. Jako liczbę stopni swobody przyjęłyśmy 10.

```r
# korzystamy z faktu, że test bowkera to test mcnemara ale dla tabel
   wiekszych niz 2x2
count <- c(10, 2, 1, 1, 0,
           0, 15, 1, 1, 0,
           1, 1, 32, 6, 0,
           0, 0, 1, 96, 3,
           1, 1, 0, 1, 26)
label <- c('-2', '-1', '0', '1', '2')
```

```r
# test bowkera
quest_table <- matrix(count, nrow = 5, byrow = TRUE)

rownames(quest_table) <- label
colnames(quest_table) <- label

mcNemar_test_with <- mcnemar.test(quest_table, correct = TRUE)
mcNemar_test_without <- mcnemar.test(quest_table, correct = FALSE)
mcNemar_test_with$p.value
mcNemar_test_without$p.value

mcNemar_test_with
mcNemar_test_without
```

```r
# test IW
quest_1 <- gl(5, 5, labels=label)
quest_2 <- gl(5, 1, labels=label)

comfort_data <- data.frame(quest_1, quest_2, count)

symmetry <- glm(count ~ Symm(quest_1, quest_2), data=comfort_data, family
   =poisson)

summary(symmetry)
```

```r
x <- symmetry$deviance # odchylenie
r <- 10 # liczba stopni swobody
p_val_IW <- 1 - pchisq(x, r)

result <- data.frame(
  "test" = c("Bowker z poprawka", "Bowker bez poprawki", "IW"),
  "p-wartość" = c(mcNemar_test_with$p.value, mcNemar_test_without$p.value
   , p_val_IW)
)
result
```

W wyniku działania powyższego kodu otrzymaliśmy p-wartości widoczne w tabeli poniżej.

test	p-wartość
Bowker z poprawka	NaN
Bowker bez poprawki	NaN
IW	0.205975

Tabela 8: P-wartości dla poszczególnych testów

Jak możemy zauważyć w tabeli 8 wynik testów Bowkera z i poprawka i bez niej daje NaN, co sugeruje, że w danych występuje problem numeryczny. Tego rodzaju problem może wynikać z obecności zerowych wartości w tabeli kontyngencji, co sprawia, że testy Bowkera nie mogą zostać prawidłowo przeprowadzone. Natomiast test IW daje p-wartość ≈ 0.205975, co sugeruje, że na poziomie istotności 0.05 nie możemy odrzucić hipotezy zerowej, zakładającej, że odpowiedzi w pierwszym i drugim okresie są symetryczne.

Aby zniwelować problem pojawiania się wartości NaN postanowiłyśmy dodać małą liczbę (równą 0.00000001) do elementów tabeli kontyngencji. Może to jednak zaburzyć wynik i tym samym mieć wpływ na poprawność otrzymanych wyników.

```
mcNemar_test_with <- mcnemar.test(quest_table + 0.00000001, correct =
    TRUE)
mcNemar_test_without <- mcnemar.test(quest_table + 0.00000001, correct =
    FALSE)
mcNemar_test_with$p.value
mcNemar_test_without$p.value
mcNemar_test_with
mcNemar_test_without

result <- data.frame(
  "test" = c("Bowker z poprawka", "Bowker bez poprawki", "IW"),
  "p-wartość" = c(mcNemar_test_with$p.value, mcNemar_test_without$p.value
    , p_val_IW)
)
result
```

test	p.wartość
Bowker z poprawka	0.391867
Bowker bez poprawki	0.391867
IW	0.205975

Tabela 9: P-wartości dla poszczególnych testów po poprawce

W tym przypadku otrzymaliśmy p-wartości widoczne w tabeli 9. Możemy w niej zauważyć, że testy Bowkera z i bez poprawki zwracają p-wartość ≈ 0.391867. Dodanie małej liczby zapobiega obecności zer w tabeli, co umożliwia przeprowadzenie testów Bowkera. Widzimy, że wyniki testów Bowkera również wskazują, że na poziomie istotności $\alpha = 0.05$ nie możemy odrzucić hipotezy zerowej. To sugeruje, że nie ma statystycznie istotnych dowodów na brak symetrii między odpowiedziami z dwóch okresów.

Podsumowując, na podstawie wyników testów Bowkera (z poprawką i bez) oraz testu IW, możemy stwierdzić, że nie ma istotnych statystycznie dowodów na to, że ocena podejścia firmy do utrzymania równowagi między życiem zawodowym, a prywatnym uległa zmianie po wdrożeniu działań mających na celu poprawę komfortu pracy. Hipoteza o symetrii odpowiedzi w obu okresach nie może być odrzucona, co sugeruje, że działania firmy nie wpłynęły znacząco na ocenę, przez respondentów, równowagi między życiem, a zawodowym.

Ogólnie możemy stwierdzić, że test ilorazu wiarygodności jest bardziej niezawodny.

5 Część III

5.1 Zadanie 6

W pewnym badaniu porównywano skuteczność dwóch metod leczenia: Leczenie A to nowa procedura, a Leczenie B to stara procedura. Przeanalizuj dane przedstawione w Tabeli 10 (wyniki dla całej grupy pacjentów) oraz w Tabelach 11 i 12 (wyniki w podgrupach ze względu na dodatkową zmienną) i odpowiedz na pytanie, czy dla danych występuje paradoks Simpsona.

Metoda	Wynik leczenia	
	Poprawa	Brak
Leczenie A	117	104
Leczenia B	177	44

Tabela 10: Dane dla całej grupy

Metoda	Reakcja	
	Poprawa	Brak
Leczenie A	17	101
Leczenia B	2	36

Tabela 11: Dane dla pacjentów z chorobami współistniejącymi

Metoda	Reakcja	
	Poprawa	Brak
Leczenie A	100	3
Leczenia B	175	8

Tabela 12: Dane dla pacjentów bez chorób współistniejących

W tym zadaniu badałyśmy zjawisko zwane paradoksem Simpsona. Występuje on, gdy tendencja zaobserwowana w kilku podgrupach znika lub odwraca się, gdy dane są łączone w jedną grupę.

```r
# Funkcja obliczająca prawdopodobieństwo poprawy
calculate_improvement_probabilities <- function(data) {
  improvement_probabilities <- numeric(nrow(data))

  for (i in 1:nrow(data)) {
    improvement_probabilities[i] <- data[i, 2] / sum(data[i, 2:3])
  }

  return(improvement_probabilities)
}
```

```r
# Funkcja wykonująca test asymptotyczny (test proporcji)
perform_prop_test <- function(data, alpha=0.05) {
  # H0 : prawdopodobieństwo pozytywnej reakcji na leczenie metodą A jest
   większe bądź równe prawdopodobieństwu pozytywnej reakcji na leczenie
   metodą B
  improvement <- c(data[1, 2], data[2, 2])
```

```
  sum_data <- c(sum(data[1, 2:3]), sum(data[2, 2:3]))

  prop_with <- prop.test(improvement,
                         sum_data,
                         alternative = "less", conf.level = 1 - alpha,
    correct = TRUE)
  prop_without <- prop.test(improvement,
                            sum_data,
                            alternative = "less", conf.level = 1 - alpha,
    correct = FALSE)

  rejection_H0_with <- prop_with$p.value < alpha
  rejection_H0_without <- prop_without$p.value < alpha

  return(list(rejection_H0_with=rejection_H0_with, rejection_H0_without=
    rejection_H0_without))
}
```

```
# Tabela 3
total_data <- data.frame(
  Methoda = c("leczenie A", "leczenie B"),
  Poprawa = c(117, 177),
  Brak = c(104, 44)
)

# Tabela 4
comorbid_data <- data.frame(
  Methoda = c("leczenie A", "leczenie B"),
  Poprawa = c(17, 2),
  Brak = c(101, 36)
)

# Tabela 5
no_comorbid_data <- data.frame(
  Methoda = c("leczenie A", "leczenie B"),
  Poprawa = c(100, 175),
  Brak = c(3, 8)
)
```

```
# Obliczanie prawdopodobieństw poprawy dla każdej tabeli
total_improvement_probabilities <- calculate_improvement_probabilities(
    total_data)
comorbid_improvement_probabilities <- calculate_improvement_probabilities
    (comorbid_data)
no_comorbid_improvement_probabilities <- calculate_improvement_
    probabilities(no_comorbid_data)

# Test proporcji (chi2 test) dla każdej tabeli
total_test_result <- perform_prop_test(total_data)
comorbid_test_result <- perform_prop_test(comorbid_data)
no_comorbid_test_result <- perform_prop_test(no_comorbid_data)
```

```
# Wyniki
results <- data.frame(
  Grupa = c("Cała grupa", "Z chorobami współistniejącymi", "Bez chorób
    współistniejących"),
  Prawdopodobieństwo_poprawy_leczenie_A = c(total_improvement_
    probabilities[1],
                                            comorbid_improvement_
    probabilities[1],
```

```
                                                no_comorbid_improvement_
    probabilities[1]),
  Prawdopodobieństwo_poprawy_leczenie_B = c(total_improvement_
    probabilities[2],
                                                comorbid_improvement_
    probabilities[2],
                                                no_comorbid_improvement_
    probabilities[2]),
  Odrzucenie_H_0_z_poprawką = c(total_test_result$rejection_H0_with,
                                comorbid_test_result$rejection_H0_with,
                                no_comorbid_test_result$rejection_H0_with
    ),
  Odrzucenie_H_0_bez_poprawki = c(total_test_result$rejection_H0_without,
                                  comorbid_test_result$rejection_H0_
    without,
                                  no_comorbid_test_result$rejection_H0_
    without)
)

results
```

W wyniku działania powyższego kodu otrzymałyśmy wyniki widoczne w poniższej tabeli.

Grupa	P-stwo poprawy leczenie A	P-stwo poprawy leczenie B	Odrzucenie H_0 z poprawką	Odrzucenie H_0 bez poprawki
Cała grupa	0.529412	0.800905	TRUE	TRUE
Z chorobami współistniejącymi	0.144068	0.052632	FALSE	FALSE
Bez chorób współistniejących	0.970874	0.956284	FALSE	FALSE

Tabela 13: Prawdopodobieństwo poprawy i odrzucenia H_0

Analizując wyniki dla całej grupy możemy stwierdzić, że leczenie B wydaje się być bardziej skuteczne niż leczenie A, ponieważ prawdopodobieństwo poprawy jest wyższe dla leczenia B (0.800905) w porównaniu do leczenia A (0.529412). Ponadto, test proporcji wskazuje na odrzucenie hipotezy zerowej (H_0) mówiącej o tym, że prawdopodobieństwo pozytywnej reakcji na leczenie metodą A jest większe bądź równe prawdopodobieństwu pozytywnej reakcji na leczenie metodą B, co sugeruje istotną statystycznie różnicę na korzyść leczenia B.

W grupie pacjentów z chorobami współistniejącymi, leczenie A wydaje się być bardziej skuteczne niż leczenie B, ponieważ prawdopodobieństwo poprawy jest wyższe dla leczenia A (0.144068) w porównaniu do leczenia B (0.052632). Test proporcji nie wskazuje istotnej statystycznie różnicy, co oznacza, że nie możemy odrzucić hipotezy zerowej (H_0).

W grupie pacjentów bez chorób współistniejących, leczenie A jest nieznacznie bardziej skuteczne niż leczenie B, ponieważ prawdopodobieństwo poprawy jest wyższe dla leczenia A (0.970874) w porównaniu do leczenia B (0.956284). Podobnie jak w poprzedniej grupie, test proporcji nie wskazuje istotnej statystycznie różnicy, więc nie możemy odrzucić hipotezy zerowej.

Możemy zauważyć, że w całej grupie leczenie B jest bardziej skuteczne niż leczenie A. Analizując poszczególne podgrupy dochodzimy do przeciwnych wniosków. Pozwala nam to

wnioskować, że mamy do czynienia z paradoksem Simpsona, ponieważ agregacja danych z obu podgrup prowadzi do przeciwnych wniosków niż analiza tych podgrup oddzielnie. Pomimo że leczenie A wydaje się być bardziej skuteczne w obu podgrupach, po połączeniu danych (cała grupa) leczenie B okazuje się bardziej skuteczne.

5.2 Zadanie 7

Dla danych z listy 1, przyjmując za zmienną 1 zmienną **CZY_KIER**, za zmienną 2 – zmienną **PYT_2** i za zmienną 3 – zmienną **STAŻ**, podaj interpretacje następujących modeli log-liniowych: [1 3], [13], [1 2 3], [12 3], [12 13] oraz [1 23].

W zadaniu zakładamy, że zmienna CZY_KIER (1) ma R możliwych odpowiedzi, zmienna PYT_2 (2) ma C możliwych odpowiedzi, a zmienna STAŻ (3) ma L możliwych odpowiedzi. Wykorzystałyśmy modele log-liniowe widoczne poniżej:

- **Model [1 3]** mówi nam o tym, że zmienne CZY_KIER (1) i STAŻ (3) mają dowolne rozkłady oraz, że zmienne CZY_KIER (1), PYT_2 (2) i STAŻ (3) są niezależne. Oznacza to, że zakładamy brak interakcji między tymi zmiennymi, a jedynie badamy, jak każda z tych zmiennych wpływa na odpowiedź.

  $$\lambda = \lambda + \lambda_i^{(1)} + \lambda_k^{(3)}, \quad \forall i \in \{1, \ldots, R\}, \quad j \in \{1, \ldots, C\}, \quad k \in \{1, \ldots, L\}$$

- **Model [13]** mówi nam o tym, że zmienne CZY_KIER (1) i STAŻ (3) mają dowolne rozkłady oraz że zmienne te nie są niezależne, zmienna PYT_2 (2) ma równomierny rozkład. Oznacza to, że badamy zarówno indywidualne wpływy tych zmiennych, jak i wpływ ich zależności.

  $$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_k^{(3)} + \lambda_{ik}^{(13)}, \quad \forall i \in \{1, \ldots, R\}, \quad j \in \{1, \ldots, C\}, \quad k \in \{1, \ldots, L\}$$

- **Model [1 2 3]** mówi nam o tym, że zmienne CZY_KIER (1), PYT_2 (2) oraz STAŻ (3) są wzajemnie niezależne. Oznacza to, że zakładamy, że każda z tych zmiennych niezależnie wpływa na odpowiedź.

  $$\lambda = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)}, \quad \forall i \in \{1, \ldots, R\}, \quad j \in \{1, \ldots, C\}, \quad k \in \{1, \ldots, L\}$$

- **Model [12 3]** mówi nam o tym, że zmienna STAŻ (3) jest niezależna od zmiennej CZY_KIER (1) i PYT_2 (2) oraz że zmienne CZY_KIER (1) i PYT_2 (2) nie są niezależne. Oznacza to, że badamy indywidualne wpływy wszystkich trzech zmiennych na odpowiedź, jak również wpływ zależności między zmiennymi CZY_KIER (1) i PYT_2 (2).

  $$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)}, \quad \forall i \in \{1, \ldots, R\}, \quad j \in \{1, \ldots, C\}, \quad k \in \{1, \ldots, L\}$$

- **Model [12 13]** mówi nam o tym, że przy ustalonej wartości zmiennej CZY_KIER (1), zmienne PYT_2 (2) oraz STAŻ (3) są niezależne. Innymi słowy zmienne PYT_2 (2) i STAŻ (3) są warunkowo niezależne. Oznacza to, że badamy indywidualne wpływy wszystkich trzech zmiennych, jak również wpływ interakcji CZY_KIER (1) z PYT_2 (2) oraz CZY_KIER (1) z STAŻ (3) na odpowiedź.

  $$l_{ijk} = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{ij}^{(12)} + \lambda_{ik}^{(13)},$$
  $$\forall i \in \{1, \ldots, R\}, \quad j \in \{1, \ldots, C\}, \quad k \in \{1, \ldots, L\}$$

- **Model [1 23]**, mówi nam o tym, że zmienna CZY_KIER (1) jest niezależna od zmiennej PYT_2 (2) i STAŻ (3) oraz że zmienne PYT_2 (2) i STAŻ (3) nie są niezależne. Oznacza to, że badamy indywidualne wpływy wszystkich trzech zmiennych, jak również wpływ zależności między zmiennymi PYT_2 (2) i STAŻ (3).

$$\lambda = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_k^{(3)} + \lambda_{jk}^{(23)}, \quad \forall i \in \{1, \ldots, R\}, \quad j \in \{1, \ldots, C\}, \quad k \in \{1, \ldots, L\}$$

6 Część IV oraz V

6.1 Zadanie 8

Przyjmując model log-liniowy [123] dla zmiennych opisanych w zadaniu 7 oszacuj prawdopobieństwa:

- że osoba pracująca na stanowisku kierowniczym jest zdecydowanie zadowolona ze swojego wynagrodzenia;
- że osoba o stażu pracy krótszym niż rok pracuje na stanowisku kierowniczym;
- że osoba o stażu pracy powyżej trzech lat nie pracuje na stanowisku kierowniczym.

Jakie byłyby oszacowania powyższych prawdopodobieństw przy założeniu modelu [12 23]?

```r
new_data <- xtabs(~ STAŻ + PYT_2 + CZY_KIER, data=data)
new_data

# dane - przedstawienie w tablicy
ftable(new_data, row.vars = c("PYT_2", "CZY_KIER"))

# liczności brzegowe
addmargins(new_data)

# zmieniamy format danych
new_data<- as.data.frame(as.table(new_data))
new_data[,-4] <- lapply(new_data[,-4], relevel, ref = 1) #ustawiamy referencje na "1"
new_data
```

```r
# MODELE
# lijk = lambda + lambda(1)i + lambda(2)j + lambda(3)k + lambda(12)ij + lambda(13)ik + lambda(23)jk + lambda(123)ijk
model_123 <- glm(Freq ~ CZY_KIER + PYT_2 + STAŻ +
                        (CZY_KIER*PYT_2) +
                        (CZY_KIER*STAŻ) +
                        (PYT_2*STAŻ) +
                        (CZY_KIER*PYT_2*STAŻ), data = new_data, family = poisson)
summary(model_123)

# lijk = lambda + lambda(1)i + lambda(2)j + lambda(3)k + lambda(12)ik + lambda(23)jk
model_12_23 <- glm(Freq ~ CZY_KIER + PYT_2 + STAŻ +
                          (CZY_KIER*PYT_2) +
                          (PYT_2*STAŻ), data = new_data, family = poisson)

```

```r
summary(model_12_23)

# dopasowanie
fitted_123 = fitted(model_123)
fitted_12_23 = fitted(model_12_23)

# porownanie
data_freq <- cbind(new_data, fitted_123, fitted_12_23)
data_freq
```

W wyniku działania powyższego kodu otrzymałyśmy tabelę widoczną poniżej. Zawiera ona wyniki dotyczące liczby obserwacji (Freq) i przewidywanych częstości (fitted) dla modeli [123] oraz [12 23]. Dane te dotyczą zmiennych STAŻ (staż pracy), PYT_2 (poziom zadowolenia z wynagrodzenia) i CZY_KIER (czy osoba pracuje na stanowisku kierowniczym).

STAŻ	PYT_2	CZY_KIER	Freq	fitted [123]	fitted [12 23]
1	-2	Nie	19	19.00	17.30
2	-2	Nie	40	40.00	38.92
3	-2	Nie	5	5.00	7.78
1	-1	Nie	3	3.00	2.70
2	-1	Nie	15	15.00	15.30
3	-1	Nie	0	$2.79 * 10^{-11}$	$4.36 * 10^{-9}$
1	1	Nie	0	$2.79 * 10^{-11}$	$2.22 * 10^{-16}$
2	1	Nie	0	$2.79 * 10^{-11}$	$2.22 * 10^{-16}$
3	1	Nie	0	$2.79 * 10^{-11}$	$1.25 * 10^{-8}$
1	2	Nie	18	18.00	15.75
2	2	Nie	68	68.00	68.25
3	2	Nie	5	5.00	7.00
1	-2	Tak	1	1.00	2.70
2	-2	Tak	5	5.00	6.08
3	-2	Tak	4	4.00	1.22
1	-1	Tak	0	$2.79 * 10^{-11}$	0.30
2	-1	Tak	2	2.00	1.70
3	-1	Tak	0	$2.79 * 10^{-11}$	$4.85 * 10^{-10}$
1	1	Tak	0	$2.79 * 10^{-11}$	$4.95 * 10^{-9}$
2	1	Tak	0	$2.79 * 10^{-11}$	$4.95 * 10^{-9}$
3	1	Tak	2	2.00	2.00
1	2	Tak	0	$2.79 * 10^{-11}$	2.25
2	2	Tak	10	10.00	9.75
3	2	Tak	3	3.00	1.00

Tabela 14: Porównanie modeli [123] i [12 23]

Analizując wartości widoczne w tabeli 14 widzimy, że w większości przypadków dopasowane liczności (fitted) dla modelu [123] są bardzo zbliżone do rzeczywistych liczności (Freq), co sugeruje, że model [123] dobrze odwzorowuje dane. Model [12 23] również dobrze odwzorowuje dane, ale w niektórych przypadkach (np. STAŻ = 3, PYT_2 = 2, CZY_KIER = Tak) istnieją większe różnice między licznościami przewidywanymi i rzeczywistymi. Tak jak zostało omówione w zadaniu 7, model [123] uwzględnia pełną interakcję między zmiennymi, przez co lepiej odzwierciedla on rzeczywiste dane niż model [12 23], który pomija niektóre

interakcje. W przypadkach, gdzie rzeczywiste liczności (Freq) są równe zero, modele również przewidują liczności bliskie zeru, co wskazuje na dobrą zgodność. Jednak model [12 23] ma pewne przypadki, w których przewidywane liczności są większe niż zero, co może wskazywać na gorsze dopasowanie modelu w tych przypadkach.

```
statistics <- data.frame(
  model = c("[123]", "[12 23]"),
  odchylenie = c(deviance(model_123), deviance(model_12_23)),
  p_wartość = c(1 - pchisq(deviance(model_123), df = df.residual(model_
    123)),
                1 - pchisq(deviance(model_12_23), df = df.residual(model_
    12_23)))
)

statistics
```

Następnie, korzystając z kodu widocznego poniżej obliczyłyśmy odchylenie standardowe i p-wartości dla modeli [123] i [12 23]. Statystyki te są widoczne w poniższej tabeli.

model	odchylenie	p_wartość
[123]	$5.021023 * 10^{-10}$	0.00
[12 23]	15.64161	0.04781

Tabela 15: Statystyki modeli [123] i [12 23]

Wiemy, że odchylenie standardowe to miara dopasowania modelu do danych. Im mniejsze odchylenie, tym lepsze dopasowanie modelu. Pozwala nam to stwierdzić, że model [123] wykazuje bardzo dobre dopasowanie do danych, ponieważ jego odchylenie standardowe jest bliskie 0. Widzimy także, że model [12 23] ma odchylenie równe ≈ 15.64161, co wskazuje na gorsze dopasowanie do danych. Potwierdza to nasze wcześniejsze przypuszczenia.

Analizując p-wartości, które oceniają czy model jest statystycznie znaczący możemy stwierdzić, że model [123] jest statystycznie istotny, ponieważ jego p-wartość jest równa $0 < 0.05 = \alpha$. W przypadku modelu [12 23] otrzymałyśmy p-wartość równą około $0.04781 < 0.05 = \alpha$ co również wskazuje na statystyczną istotność. Wartość ta jest jednak bliska poziomowi istotności α.

W kolejnym kroku zajęłyśmy się oszacowaniem poszczególnych prawdopodobieństw.

```
# 1. Osoba pracująca na stanowisku kierowniczym jest zdecydowanie
    zadowolona ze swojego wynagrodzenia
# P(PYT_2 = 2 | CZY_KIER = Tak) = P(PYT_2 = 2 i CZY_KIER = Tak)/P(CZY_
    KIER = Tak)

prob1_numerator <- sum(data_freq$Freq[data_freq$CZY_KIER == "Tak" & data_
    freq$PYT_2 == "2"])
prob1_denominator <- sum(data_freq$Freq[data_freq$CZY_KIER == "Tak"])

prob1 <- prob1_numerator / prob1_denominator

prob1_model_123_numerator <- sum(data_freq$fitted_123[data_freq$CZY_KIER
    == "Tak" & data_freq$PYT_2 == "2"])
prob1_model_123_denominator <- sum(data_freq$fitted_123[data_freq$CZY_
    KIER == "Tak"])

model_123_prob1 <- prob1_model_123_numerator/prob1_model_123_denominator

prob1_model_12_23_numerator <- sum(data_freq$fitted_12_23[data_freq$CZY_
    KIER == "Tak" & data_freq$PYT_2 == "2"])
```

```r
prob1_model_12_23_denominator <- sum(data_freq$fitted_12_23[data_freq$CZY_KIER == "Tak"])

model_12_23_prob1 <- prob1_model_12_23_numerator/prob1_model_12_23_denominator


# 2. Osoba o stażu pracy krótszym niż rok pracuje na stanowisku kierowniczym
#  P(CZY_KIER = Tak | STAŻ = 1) = P(CZY_KIER = Tak i STAŻ = 1)/P(STAŻ = 1)

prob2_numerator <- sum(data_freq$Freq[data_freq$CZY_KIER == "Tak" & data_freq$STAŻ == "1"])
prob2_denominator <- sum(data_freq$Freq[data_freq$STAŻ == "1"])

prob2 <- prob2_numerator / prob2_denominator

prob2_model_123_numerator <- sum(data_freq$fitted_123[data_freq$CZY_KIER == "Tak" & data_freq$STAŻ == "1"])
prob2_model_123_denominator <- sum(data_freq$fitted_123[data_freq$STAŻ == "1"])

model_123_prob2 <- prob2_model_123_numerator/prob2_model_123_denominator

prob2_model_12_23_numerator <- sum(data_freq$fitted_12_23[data_freq$CZY_KIER == "Tak" & data_freq$STAŻ == "1"])
prob2_model_12_23_denominator <- sum(data_freq$fitted_12_23[data_freq$STAŻ == "1"])

model_12_23_prob2 <- prob2_model_12_23_numerator/prob2_model_12_23_denominator


# 3. Osoba o stażu pracy powyżej trzech lat nie pracuje na stanowisku kierowniczym
# P(CZY_KIER = Nie | STAŻ = 3) = P(CZY_KIER = Nie i STAŻ = 3)/P(STAŻ = 3)

prob3_numerator <- sum(data_freq$Freq[data_freq$CZY_KIER == "Nie" & data_freq$STAŻ == "3"])
prob3_denominator <- sum(data_freq$Freq[data_freq$STAŻ == "3"])

prob3 <- prob3_numerator / prob3_denominator

prob3_model_123_numerator <- sum(data_freq$fitted_123[data_freq$CZY_KIER == "Nie" & data_freq$STAŻ == "3"])
prob3_model_123_denominator <- sum(data_freq$fitted_123[data_freq$STAŻ == "3"])

model_123_prob3 <- prob3_model_123_numerator/prob3_model_123_denominator

prob3_model_12_23_numerator <- sum(data_freq$fitted_12_23[data_freq$CZY_KIER == "Nie" & data_freq$STAŻ == "3"])
prob3_model_12_23_denominator <- sum(data_freq$fitted_12_23[data_freq$STAŻ == "3"])

model_12_23_prob3 <- prob3_model_12_23_numerator/prob3_model_12_23_denominator
```

```r
# wyniki prawdopodobienstw
results_summary <- data.frame(
  Scenariusz = c("P(PYT_2 = 2 | CZY_KIER = Tak)", "P(CZY_KIER = Tak | STAŻ = 1)", "P(CZY_KIER = Nie | STAŻ = 3)"),
  Z_danych = c(prob1, prob2, prob3),
  Model_123 = c(model_123_prob1, model_123_prob2, model_123_prob3),
  Model_12_23 = c(model_12_23_prob1, model_12_23_prob2, model_12_23_prob3)
)

results_summary
xtable(results_summary)
```

Scenariusz	Z danych	Model [123]	Model [12 23]
P(PYT_2 = 2 | CZY_KIER = Tak)	0.481482	0.481482	0. 0.481482
P(CZY_KIER = Tak | STAŻ = 1)	0.024390	0.024390	0.128115
P(CZY_KIER = Nie | STAŻ = 3)	0.526316	0.526315	0.778094

Tabela 16: Wartości oszacowanych prawdopodobieństw dla obu modeli

W wyniku działania kodu widocznego powyżej otrzymałyśmy tabelkę 16, w której widoczne są wartości oszacowanych prawdopodobieństw dla modeli [123] oraz [12 23] dla każdego z 3 podpunktów.

W podpunkcie 1. badaliśmy prawdopodobieństwo, że osoba pracująca na stanowisku kierowniczym jest zdecydowanie zadowolona ze swojego wynagrodzenia. Analizując wartości widoczne w tabeli 16 widzimy, że zarówno wartość z danych, jak i oba modele oszacowują to prawdopodobieństwo na 0.481482. Oznacza to, że około 48.1% osób na stanowiskach kierowniczych jest zdecydowanie zadowolonych ze swojego wynagrodzenia, niezależnie od przyjętego modelu.

W podpunkcie 2. badaliśmy prawdopodobieństwo, że osoba o stażu pracy krótszym niż rok pracuje na stanowisku kierowniczym. Analizując wartości widoczne w tabeli 16 widzimy, że wartość z danych wynosi 0.024390, co oznacza, że około 2.4% osób o stażu pracy krótszym niż rok pracuje na stanowisku kierowniczym. Model [123] oszacowuje to prawdopodobieństwo na 0.024390, co jest zgodne z danymi. Model [12 23] oszacowuje to prawdopodobieństwo na 0.128115, co jest znacznie wyższe niż wartość z danych. Potwierdza nam to wcześniejszą tezę mówiącą o tym, że model [123] lepiej odwzorowuje dane niż model [12 23].

Natomiast w podpunkcie 3. badaliśmy prawdopodobieństwo, że osoba o stażu pracy powyżej trzech lat nie pracuje na stanowisku kierowniczym. Analizując wartości widoczne w tabeli 16 widzimy, że wartość z danych wynosi 0.526316, co oznacza, że około 52.6% osób o stażu pracy powyżej trzech lat nie pracuje na stanowisku kierowniczym. Model [123] oszacowuje to prawdopodobieństwo na 0.526315, co jest zgodne z danymi. Model [12 23] oszacowuje to prawdopodobieństwo na 0.778094, co jest znacznie wyższe niż wartość z danych. Potwierdza to naszą tezę.

Podsumowując, model [123] jest znacznie lepiej dopasowany do danych niż model [12 23], co sugeruje, że powinien być preferowany przy analizie zależności między zmiennymi. Model [123] dokładnie odwzorowuje prawdopodobieństwa obliczone z danych dla wszystkich trzech scenariuszy. Model [12 23] znacznie różni się w oszacowaniach prawdopodobieństw dla scenariuszy związanych z zależnością między stażem pracy a stanowiskiem kierowniczym, co sugeruje, że ten model może nie uwzględniać pewnych istotnych interakcji między zmiennymi.

6.2 Zadanie 9

Dla danych wskazanych w zadaniu 7 zweryfikuj następujące hipotezy:

- zmienne losowe **CZY_KIER**, **PYT_2** i **STAŻ** są wzajemnie niezależne;
- zmienna losowa **PYT_2** jest niezależna od pary zmiennych **CZY_KIER** i **STAŻ**;
- zmienna losowa **PYT_2** jest niezależna od zmiennej **CZY_KIER**, przy ustalonej wartości zmiennej **STAŻ**

W tym zadaniu przyjęłyśmy poziom istotności $\alpha = 0.05$.

```
# zmienne losowe CZY_KIER, PYT_2 i STAZ sa wzajemnie niezalezne;
# H_0: M_0 = [1 2 3]
model_1_2_3 <- glm(Freq ~ CZY_KIER + PYT_2 + STAŻ, data = new_data,
    family = poisson)

# H_1: M = [123]
model_123 <- glm(Freq ~ CZY_KIER + PYT_2 + STAŻ +
                   (CZY_KIER*PYT_2) +
                   (CZY_KIER*STAŻ) +
                   (PYT_2*STAŻ) +
                   (CZY_KIER*PYT_2*STAŻ), data = new_data, family =
    poisson)
anova_test_1 <- anova(model_1_2_3, model_123)
anova_p_val_1 <- 1 - pchisq(anova_test_1$Deviance[2], df = anova_test_1$
    Df[2])
anova_p_val_1
# p_val < 0.05 wiec odrzucamy H_0

# H_1: M = [12 23 31]
model_12_23_31 <- glm(Freq ~ (CZY_KIER + PYT_2 + STAŻ)^2, data = new_data
    , family = poisson)
anova_test_2 <- anova(model_1_2_3, model_12_23_31)
anova_p_val_2 <- 1 - pchisq(anova_test_2$Deviance[2], df = anova_test_2$
    Df[2])
anova_p_val_2
# p_val < 0.05 wiec odrzucamy H_0

# WYNIKI
table1 <- data.frame(
  Hipoteza = "CZY_KIER, PYT_2 i STAZ są wzajemnie niezależne",
  Model = c("[123]", "[12 23 31]"),
  p_value = c(anova_p_val_1, anova_p_val_2),
  Decyzja = c(ifelse(anova_p_val_1 < 0.05, "Odrzucamy H_0", "Nie
    odrzucamy H_0"),
              ifelse(anova_p_val_2 < 0.05, "Odrzucamy H_0", "Nie
    odrzucamy H_0"))
)
table1
```

Hipoteza	Model	p-wartość	Decyzja
CZY_KIER, PYT_2 i STAŻ są wzajemnie niezależne	[123]	$6.19 * 10^{-4}$	Odrzucamy H_0
CZY_KIER, PYT_2 i STAŻ są wzajemnie niezależne	[12 23 31]	$2.77 * 10^{-5}$	Odrzucamy H_0

Tabela 17: Parametry do weryfikacji hipotezy - podpunkt 1

W pierwszej kolejności chciałyśmy zweryfikować hipotezę zerową mówiącą o tym, że zmienne losowe CZY_KIER (1), PYT_2 (2) i STAŻ (3) są wzajemnie niezależne. Jak możemy zauważyć w tabelce 17 została ona odrzucona we wszystkich modelach. Oznacza, że istnieją istotne interakcje między tymi zmiennymi. Zatem, te zmienne nie są wzajemnie niezależne.

```r
# zmienna losowa PYT_2(2) jest niezalezna od pary zmiennych CZY_KIER(1) i
    STAZ(3);

# H_0: M_0 = [12 3]
model_1_23 <- glm(Freq ~ CZY_KIER + PYT_2 + STAŻ +
                    (CZY_KIER*STAŻ), data = new_data, family = poisson)

# H_1: M = [123]
anova_test_3 <- anova(model_1_23, model_123)
anova_p_val_3 <- 1 - pchisq(anova_test_3$Deviance[2], df = anova_test_3$
    Df[2])
anova_p_val_3
# p_val > 0.05 wiec nie mamy podstaw do odrzucenia H_0

# H_1: M = [12 23 31]
anova_test_4 <- anova(model_1_23, model_12_23_31)
anova_p_val_4 <- 1 - pchisq(anova_test_4$Deviance[2], df = anova_test_4$
    Df[2])
anova_p_val_4
# p-val < 0.05, odrzucamy H_0

# WYNIKI
table2 <- data.frame(
  Hipoteza = "PYT_2 jest niezależna od pary CZY_KIER i STAZ",
  Model = c("[123]", "[12 23 31]"),
  p_value = c(anova_p_val_3, anova_p_val_4),
  Decyzja = c(ifelse(anova_p_val_3 < 0.05, "Odrzucamy H_0", "Nie
    odrzucamy H_0"),
              ifelse(anova_p_val_4 < 0.05, "Odrzucamy H_0", "Nie
    odrzucamy H_0"))
)

table2
```

Hipoteza	Model	p-wartość	Decyzja
PYT_2 jest niezależna od pary CZY_KIER i STAŻ	[123]	0.080963	Nie odrzucamy H_0
PYT_2 jest niezależna od pary CZY_KIER i STAŻ	[12 23 31]	0.010400	Odrzucamy H_0

Tabela 18: Parametry do weryfikacji hipotezy - podpunkt 2

Następnie zweryfikowałyśmy hipotezę zerową mówiącą o tym, że zmienna losowa PYT_2 (2) jest niezależna od pary zmiennych CZY_KIER (1) i STAŻ (3). W tabeli 18 możemy zauważyć, że w przypadku modelu [123], hipoteza zerowa nie została odrzucona (p-wartość $= 0.08 > 0.05$), co sugeruje, że zmienna PYT_2 jest niezależna od pary CZY_KIER i STAŻ w tym modelu. Jednak w modelu [12 23 31], hipoteza została odrzucona (p-wartość $= 0.01 < 0.05$), co sugeruje, że w tym modelu istnieją istotne interakcje i zależności między zmiennymi.

```
# zmienna losowa PYT_2 jest niezalezna od zmiennej CZY_KIER, przy ustalonej wartosci zmiennej STAZ.

# H_0: M_0 = [12 23]
model_12_23 <- glm(Freq ~ CZY_KIER + PYT_2 + STAŻ +
                     (CZY_KIER*STAŻ) +
                     (PYT_2*STAŻ), data = new_data, family = poisson)


# H_1: M = [123]
p_val_1 <- 1 - pchisq(deviance(model_12_23) - deviance(model_123),
                      df = df.residual(model_12_23) - df.residual(model_123))
p_val_1

# p_val > 0.05 wiec nie mamy podstaw do odrzucenia H_0

# H_1: M = [12 23 31]
p_val_2 <- 1 - pchisq(deviance(model_12_23) - deviance(model_12_23_31),
                      df = df.residual(model_12_23) - df.residual(model_12_23_31))
p_val_2

# p_val > 0.05 wiec nie mamy podstaw do odrzucenia H_0


# WYNIKI
table3 <- data.frame(
  Hipoteza = "PYT_2 jest niezależna od CZY_KIER przy ustalonej wartości STAZ",
  Model = c("[123]", "[12 23 31]"),
  p_value = c(p_val_1, p_val_2),
  Decyzja = c(ifelse(p_val_1 < 0.05, "Odrzucamy H_0", "Nie odrzucamy H_0"),
              ifelse(p_val_2 < 0.05, "Odrzucamy H_0", "Nie odrzucamy H_0"))
)

table3
```

Hipoteza	Model	p-wartość	Decyzja
PYT_2 jest niezależna od CZY_KIER przy ustalonej wartości STAŻ	[123]	0.844645	Nie odrzucamy H_0
PYT_2 jest niezależna od CZY_KIER przy ustalonej wartości STAŻ	[12 23 31]	0.349985	Nie odrzucamy H_0

Tabela 19: Parametry do weryfikacji hipotezy - podpunkt 3

Zweryfikowałyśmy także hipotezę zerową mówiącą o tym, że zmienna losowa PYT_2 (2) jest niezależna od zmiennej CZY_KIER (1), przy ustalonej wartości zmiennej STAŻ (3). W tabeli 19 widzimy, że w obu modelach, hipoteza zerowa nie została odrzucona (p-wartość > 0.05), co sugeruje, że zmienna PYT_2 jest niezależna od zmiennej CZY_KIER, gdy zmienna STAŻ jest ustalona. Oznacza to, że przy stałej wartości STAŻ, zmienne PYT_2 i CZY_KIER nie mają istotnych interakcji.

7 Źródła

- https://www.r-project.org/other-docs.html
- https://www.rdocumentation.org/