Sprawozdanie 2 ADA

Krzysztof Radomski 275968

10 grudnia 2024

Spis treści

1 Zadanie 1

```
tabela <- matrix(c(4,0,3,17),nrow=2,ncol=2)
zad_1_f <- fisher.test(tabela)$p.value
cat("p-value dla testu Fishera, dla danych z zadania to:", zad_1_f, "\n")

## p-value dla testu Fishera, dla danych z zadania to: 0.003293808

zad_1_b <- boschloo(x1=4,x2=3,n1=4,n2=20)$p.value
cat("p-value dla testu Boschloo, dla danych z zadania to:", zad_1_b, "\n")

## p-value dla testu Boschloo, dla danych z zadania to: 0.002055041
```

Dane z zadania ułożyłem w tabelkę 2x2, i zastosowałem test Fishera oraz Boschloo. Na podstawie obu p-wartości możemy stwierdzić, że badane zmienne nie są niezależne, bo w obu przypadkach odrzuciliśmy hipotezę H_0 o niezależności na poziomie istotności $\alpha = 0.05$. Z zadania 3 dowiemy się o tym, że test Boschloo'ego jest lepszy dlatego ostatecznie przyjmujemy, że nasza p-wartość wynosi 0.002055

2 Zadanie 2

```
#a)
tabelka_a <- table(new_data$female, new_data$wageCat)
fisher.test(tabelka_a, simulate.p.value = TRUE)
p_fh_a <- fisher.test(tabelka_a)$p.value

#b)
tabelka_b <- table(new_data$married, new_data$wageCat)
p_fh_b <- fisher.test(tabelka_b, simulate.p.value = TRUE)$p.value

#c)
tabelka_c <- table(new_data$region, new_data$wageCat)
p_fh_c <- fisher.test(tabelka_c, simulate.p.value = TRUE)$p.value

#d)
tabelka_d <- table(new_data$educCat, new_data$wageCat)
p_fh_d <- fisher.test(tabelka_d, simulate.p.value = TRUE)$p.value

#e)
tabelka_e <- table(new_data$experCat, new_data$wageCat)
p_fh_e <- fisher.test(tabelka_e, simulate.p.value = TRUE)$p.value

#f)
tabelka_f <- table(new_data$educCat2, new_data$wageCat)
p_fh_f <- fisher.test(tabelka_f, simulate.p.value = TRUE)$p.value
```

Przeprowadzono dokładny test Fishera dla zmiennych kategorycznych z zestawu danych `new_data` oraz zmiennej `wageCat` (kategorie wynagrodzeń) na poziomie istotności $\alpha = 0.05$. Wyniki testów oraz wnioski przedstawiono poniżej:

1. **Porównanie zmiennej `female` z `wageCat`:**
   - Wynik z symulowaną p-wartością: $p = 0.0004998$.
   - Wynik bez symulacji: $p = 2.2 \times 10^{-16}$.
   - Odrzucamy hipotezę o niezależności płci od kategorii wynagrodzenia.
   - **Wniosek**: Płeć jest statystycznie istotnie związana z kategoriami wynagrodzenia.
2. **Porównanie zmiennej `married` z `wageCat`:**
   - Wynik z symulowaną p-wartością: $p = 0.0004998$.
   - Odrzucamy hipotezę o niezależności stanu cywilnego od kategorii wynagrodzenia.
   - **Wniosek**: Stan cywilny jest statystycznie istotnie związany z kategoriami wynagrodzenia.
3. **Porównanie zmiennej `region` z `wageCat`:**
   - Wynik z symulowaną p-wartością: $p = 0.4343$.
   - Brak podstaw do odrzucenia hipotezy o niezależności regionu od kategorii wynagrodzenia.
   - **Wniosek**: Nie stwierdzono statystycznie istotnego związku między regionem zamieszkania a kategoriami wynagrodzenia.
4. **Porównanie zmiennej `educCat` z `wageCat`:**
   - Wynik z symulowaną p-wartością: $p = 0.0004998$.
   - Odrzucamy hipotezę o niezależności wykształcenia od kategorii wynagrodzenia.
   - **Wniosek**: Wykształcenie jest statystycznie istotnie związane z kategoriami wynagrodzenia.
5. **Porównanie zmiennej `experCat` z `wageCat`:**
   - Wynik z symulowaną p-wartością: $p = 0.0004998$.
   - Odrzucamy hipotezę o niezależności doświadczenia zawodowego od kategorii wynagrodzenia.
   - **Wniosek**: Doświadczenie zawodowe jest statystycznie istotnie związane z kategoriami wynagrodzenia.
6. **Porównanie zmiennej `educCat2` z `wageCat`:**
   - Wynik z symulowaną p-wartością: $p = 0.0004998$.
   - Odrzucamy hipotezę o niezależności alternatywnych kategorii wykształcenia od kategorii wynagrodzenia.
   - **Wniosek**: Alternatywne podziały wykształcenia również wskazują na statystycznie istotny związek z kategoriami wynagrodzenia.

W podpunkcie (a), dla zmiennej `female`, wyniki obliczono obiema metodami:

- Symulowana p-wartość: $p = 0.0004998$.
- Dokładna p-wartość: $p = 2.2 \times 10^{-16}$.

Różnica w wynikach wynika z faktu, że metoda symulowana bazuje na ograniczonej liczbie replikacji (2000), co może prowadzić do mniej dokładnego oszacowania w przypadku małych p-wartości. Dokładna metoda w tej sytuacji jest bardziej wiarygodna, ponieważ wykorzystuje wszystkie możliwe tablice kontyngencji, eliminując wpływ losowości.

Dlaczego symulowano p-wartości w innych przypadkach?

W przypadku większych tablic kontyngencji (np. `educCat`, `experCat`), liczba możliwych kombinacji wyników jest zbyt duża, aby metoda dokładna była możliwa do zastosowania. Dlatego w tych analizach zastosowano metodę symulowaną.

Wniosek

Gdy dostępna jest dokładna p-wartość (jak w podpunkcie (a)), należy ją uznać za bardziej wiarygodną. W pozostałych przypadkach metoda symulowana jest akceptowalnym przybliżeniem, o ile liczba replikacji jest odpowiednio duża. W naszym przypadku, liczba 2000 replikacji powinna być wystarczająca do uzyskania wiarygodnych wyników, choć nadal należy interpretować je z pewną ostrożnością.

3 Zadanie 3

Celem analizy jest porównanie mocy dwóch testów statystycznych: testu Fishera i testu Boschloo'ego. Moc testu to prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest fałszywa, czyli zdolność testu do wykrywania rzeczywistego efektu. Dla każdego rozmiaru próby i zestawu parametrów symulacja powtarzana jest M = 100 razy. Na koniec każdej symulacji dowiadujemy się jaką skuteczność miał test Fishera i Boschloo.

```
symulacja <- function(n_1, n_2, p_1, p_2, alpha = 0.05, M = 100){
  fisher_sum <- 0
  boschlo_sum <- 0

  for (i in 1:M) {
    x_1 <- rbinom(1, n_1, p_1)
    x_2 <- rbinom(1, n_2, p_2)
    tabela <- matrix(c(x_1, x_2, n_1 - x_1, n_2 - x_2), ncol = 2, nrow = 2)

    fisher_pval <- fisher.test(tabela)$p.value
    boschloo_pval <- boschloo(x1 = x_1, x2 = x_2, n1 = n_1, n2 = n_2,
                              alternative = "two.sided")$p.value


    # Liczba przypadków, w których p-value jest mniejsze niż alpha
    if (fisher_pval < alpha) fisher_sum <- fisher_sum + 1
```

```
    if (boschloo_pval < alpha) boschlo_sum <- boschlo_sum + 1

    return(c(fisher_sum, boschlo_sum) / M)  # Zwrot wyników
    }

  }
```

Tabela 1: Porównanie wyników testów dla próby o rozmiarze 30.

Prawdopodobienstwa	Fisher	Boschloo
p1=0.5, p2=0.5	0.01	0.02
p1=0.8, p2=0.8	0.02	0.05
p1=0.3, p2=0.4	0.11	0.17
p1=0.5, p2=0.8	0.64	0.73

Tabela 2: Porównanie wyników testów dla próby o rozmiarze 50.

Prawdopodobienstwa	Fisher	Boschloo
p1=0.5, p2=0.5	0.03	0.04
p1=0.8, p2=0.8	0.06	0.06
p1=0.3, p2=0.4	0.10	0.11
p1=0.5, p2=0.8	0.86	0.92

Tabela 3: Porównanie wyników testów dla próby o rozmiarze 100.

Prawdopodobienstwa	Fisher	Boschloo
p1=0.5, p2=0.5	0.04	0.04
p1=0.8, p2=0.8	0.04	0.06
p1=0.3, p2=0.4	0.30	0.34
p1=0.5, p2=0.8	1.00	1.00

1. **Porównanie mocy testów**: - Test Boschloo'ego lepiej radzi sobie w większości przypadków, szczególnie przy małych próbkach. - Dla konfiguracji z większymi różnicami między p_1 i p_2 (np. $p_1 = 0.5, p_2 = 0.8$), Boschloo osiąga wyższą moc niż test Fishera.

2. **Dlaczego Boschloo jest lepszy?**: - Test Boschloo'ego jest rozszerzeniem testu Fishera, które minimalizuje nadmierną konserwatywność i lepiej wykorzystuje dostępne dane, zwłaszcza przy małych próbkach.

3. **Czy Boschloo zawsze jest lepszy?**: - Nie. W przypadku dużych próbek i małych różnic między p_1 i p_2, różnica w mocy jest niewielka, co czyni test Fishera wystarczającym w takich sytuacjach.

4 Zadanie 4

```r
#a)
p_chi_a <- chisq.test(tabelka_a, correct = TRUE)$p.value
p_chi_a

## [1] 3.233533e-16

#b)
p_chi_b <- chisq.test(tabelka_b, correct = TRUE)$p.value
p_chi_b

## [1] 7.049861e-08

#c)
p_chi_c <- chisq.test(tabelka_c, correct = TRUE)$p.value
p_chi_c

## [1] 0.4446242

#d)
p_chi_d <- chisq.test(tabelka_d, correct = TRUE)$p.value
p_chi_d

## [1] 1.603004e-12

#e)
p_chi_e <- chisq.test(tabelka_e, correct = TRUE)$p.value
p_chi_e

## [1] 5.443874e-06

#f)
p_chi_f <- chisq.test(tabelka_f, correct = TRUE)$p.value
p_chi_f

## [1] 8.245634e-18
```

Podobnie jak w zadaniu 2-gim użyliśmy testów do sprawdzenia niezależności dwóch zmiennych. Dane którymi się posłużyliśmy są tymi samymi co w zadaniu 2, sprawdzamy więc niezależność tych samych par zmiennych, jednak w tym przypadku użyliśmy testu χ^2 - Pearsona

- **(a) wageCat a female:** Test niezależności wskazał, że istnieje statystycznie istotny związek pomiędzy zmiennymi `wageCat` i `female`, ponieważ wartość p wyniosła 3.23×10^{-16}, co jest znacznie mniejsze niż przyjęty poziom istotności $\alpha = 0.05$, zatem można odrzucić hipotezę zerową
- **(b) wageCat a married:** Wartość p testu wyniosła 7.05×10^{-8}, co oznacza, że zmienne `wageCat` i `married` również nie są niezależne przy poziomie istotności $\alpha = 0.05$.
- **(c) wageCat a region:** W tym przypadku wartość p wyniosła 0.4446, więc nie ma podstaw do odrzucenia hipotezy zerowej, co sugeruje brak istotnego związku pomiędzy `wageCat` a `region`.

- **(d) wageCat a educCat:** Wartość p wyniosła 1.60×10^{-12}, co wskazuje na statystycznie istotny związek między zmiennymi `wageCat` i `educCat`.
- **(e) wageCat a experCat:** Test niezależności wskazał wartość $p = 5.44 \times 10^{-6}$, co oznacza, że zmienne `wageCat` i `experCat` nie są niezależne.
- **(f) wageCat a educCat2:** Otrzymano wartość $p = 8.25 \times 10^{-18}$, co sugeruje istnienie bardzo silnego związku pomiędzy `wageCat` i zmienną `educCat2`.

Podsumowując, testy wskazują na zależność zmiennej `wageCat` od większości badanych zmiennych, z wyjątkiem zmiennej `region`. Wyniki te mogą być pomocne w dalszej analizie struktury danych i modelowaniu.

5 Zadanie 5

```
test_iw <- function(x, alpha = 0.05) {
  if (!is.matrix(x)) stop("Dane wejściowe muszą być macierzą.")

  # Suma całkowita
  n <- sum(x)
  # Liczba wierszy i kolumn
  r <- dim(x)[1]
  c <- dim(x)[2]

  # Suma brzegowa
  n_plus_j <- colSums(x)
  n_i_plus <- rowSums(x)

  # Statystyka lambda
  lambda <- 0
  for (i in 1:r) {
    for (j in 1:c) {
      expected <- (n_i_plus[i] * n_plus_j[j]) / n  # Wartości oczekiwane
      if (expected > 0) {  # Unikaj problemów z logarytmami
        lambda <- lambda + x[i, j] * log(x[i, j] / expected)
      }
    }
  }

  # Obliczanie statystyki testowej
  test_statistic <- 2 * lambda
  df <- (r - 1) * (c - 1)  # Liczba stopni swobody
  p_value <- 1 - pchisq(test_statistic, df)

  return(list(statistic = test_statistic, p_value = p_value))
}
```

```r
# Obliczanie p-value za pomocą test_iw
p_iw_a <- test_iw(as.matrix(tabelka_a))$p_value
p_iw_b <- test_iw(as.matrix(tabelka_b))$p_value
p_iw_c <- test_iw(as.matrix(tabelka_c))$p_value
p_iw_d <- test_iw(as.matrix(tabelka_d))$p_value
p_iw_e <- test_iw(as.matrix(tabelka_e))$p_value
p_iw_f <- test_iw(as.matrix(tabelka_f))$p_value
```

Poniżej przedstawiono wyniki analizy p-value dla różnych zmiennych w stosunku do zmiennej `wageCat`. Wyniki uzyskano za pomocą funkcji `test_iw`.

1. P-value dla tabelki **a** (`female vs wageCat`):

$$1.110223 \times 10^{-16}$$

   Wynik wskazuje na zależność między zmiennymi, ponieważ p-value jest mniejsze niż przyjęty poziom istotności (< 0.05).

2. P-value dla tabelki **b** (`married vs wageCat`): p-wartość wyniosła

$$6.718655 \times 10^{-8}$$

   Zależność również jest istotna statystycznie, co sugeruje różnice w kategorii `wageCat` w zależności od zmiennej `married`.

3. P-value dla tabelki **c** (`region vs wageCat`): p-wartość wyniosła

$$0.4368911$$

   zatem nie ma podstaw do odrzucenia hipotezy zerowej o niezależności między `region` a kategorią wynagrodzenia.

4. P-value dla tabelki **d** (`educCat vs wageCat`):

$$2.2315483 \times 10^{-13}, 0.05$$

   Wynik wskazuje na istotną zależność, co jest zgodne z intuicyjnym oczekiwaniem, że poziom wykształcenia wpływa na kategorię wynagrodzenia.

5. P-value dla tabelki **e** (`experCat vs wageCat`):

$$2.6920625 \times 10^{-6}$$

   Wynik potwierdza istotną zależność ($p < 0.05$), doświadczenie w wykonywanym zawodzie ma wpływ na wielkość zarobków

6. P-value dla tabelki **f** (`educCat2 vs wageCat`):

$$0$$

   Jest on równy zero zapewne z powodu tego, że p-wartością była tak mała liczba, że komputer uznał ją za zero. Wynik zatem wskazuje na istotną zależność, co oznacza, że poziom wykształcenia jest powiązany z późniejszymi zarobkami .

Wnioski

- Istotne statystycznie zależności ($p < 0.05$) zaobserwowano dla zmiennych: `female`, `married`, `educCat`, `experCat`, i `educCat2`. Oznacza to, że te zmienne są silnie powiązane z kategorią wynagrodzenia (`wageCat`).
- Brak istotnej zależności stwierdzono jedynie w przypadku zmiennej `region`. Można przypuszczać, że miejsce zamieszkania nie ma bezpośredniego wpływu na kategorię wynagrodzenia w analizowanym zbiorze danych.

6 Zadanie 6

Tabela 4: Podsumowanie wartości p dla testów niezależności (Freeman-Halton, chi-kwadrat, iloraz wiarogodności)

Test	Freeman-Halton	Chi-squared	Likelihood Ratio
A (female vs wageCat)	7.545727e-17	3.233533e-16	1.110223e-16
B (married vs wageCat)	4.997501e-04	7.049861e-08	6.718655e-08
C (region vs wageCat)	4.607696e-01	4.446242e-01	4.368911e-01
D (educCat vs wageCat)	4.997501e-04	1.603004e-12	2.231548e-13
E (experCat vs wageCat)	4.997501e-04	5.443874e-06	2.692062e-06
F (educCat2 vs wageCat)	4.997501e-04	8.245634e-18	0.000000e+00

W celu oceny zależności pomiędzy zmiennymi w zbiorze danych, przeprowadzono trzy testy statystyczne: test Freeman-Halton, test Chi-kwadrat oraz test Ilorazu Wiarogodności. Poniżej przedstawiono analizę wyników uzyskanych dla różnych par zmiennych.

- **Test Freeman-Halton:** Dla większości przypadków (A, B, D, E, F), p-wartości są bardzo małe, co sugeruje silną zależność zmiennych. Test Freeman-Halton, będący testem dokładnym, jest szczególnie przydatny w przypadku małych prób i tabel z niewielkimi licznościami, gdzie inne testy mogą dawać mniej dokładne wyniki.
- **Test Chi-kwadrat:** P-wartości dla testu Chi-kwadrat są podobne do tych uzyskanych w teście Freeman-Halton, ale w przypadku zmiennych z małymi licznościami, test Chi-kwadrat może być mniej wiarygodny. Dla przypadku C (region vs wageCat) p-wartość wyniosła 0.446, co wskazuje na brak zależności między zmiennymi, jednakże małe liczności w tabeli mogą wpłynąć na wiarygodność tego wyniku.
- **Test Ilorazu Wiarogodności:** Wyniki p dla testu Ilorazu Wiarogodności są zbliżone do tych uzyskanych w teście Chi-kwadrat, co sugeruje, że dla prostych tabel kontyngencji obie metody są równoważne. Test Ilorazu Wiarogodności jest bardziej elastyczny i może być bardziej odpowiedni w bardziej złożonych modelach.

7 Zadanie 7

7.1 Miara τ

Miara τ opisuje siłę zależności między zmiennymi w tabeli kontyngencji. Jest definiowana wzorem:

$$\tau = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}^2}{nn_{i+}} - \sum_{j=1}^{C}\left(\frac{n_{+j}}{n}\right)^2}{1 - \sum_{j=1}^{C}\left(\frac{n_{+j}}{n}\right)^2}$$

gdzie:

- n_{ij} to liczba elementów w i-tym wierszu i j-tej kolumnie,
- n_{i+} to suma elementów w i-tym wierszu,
- n_{+j} to suma elementów w j-tej kolumnie,
- n to całkowita liczba elementów w tabeli.

```
tau <- function(x) {
  total_count <- sum(x)
  num_rows <- dim(x)[1]
  num_cols <- dim(x)[2]
  col_totals <- numeric(num_cols)
  row_totals <- numeric(num_rows)

  for (col_idx in 1:num_cols) col_totals[col_idx] <- sum(x[, col_idx])
  for (row_idx in 1:num_rows) row_totals[row_idx] <- sum(x[row_idx, ])

  numerator_part <- 0
  denominator_part <- 0

  for (col_idx in 1:num_cols) {
    for (row_idx in 1:num_rows) {
      numerator_part <- numerator_part + x[row_idx, col_idx]^2 / total_count / row_total
    }
    denominator_part <- denominator_part + (col_totals[col_idx] / total_count)^2
  }

  return((numerator_part - denominator_part) / (1 - denominator_part))
}
```

7.2 Miara γ

Miara γ oblicza różnicę między liczbą par zgodnych (C) i niezgodnych (D) względem ich sumy:

$$\gamma = \frac{C - D}{C + D}$$

gdzie:

$$C = \sum_{i=1}^{R-1} \sum_{j=1}^{C-1} n_{ij} \cdot \sum_{k=i+1}^{R} \sum_{l=j+1}^{C} n_{kl}$$

to liczba par zgodnych, a

$$D = \sum_{i=2}^{R} \sum_{j=1}^{C-1} n_{ij} \cdot \sum_{k=1}^{i-1} \sum_{l=j+1}^{C} n_{kl}$$

to liczba par niezgodnych.

```
gamma <- function(x) {
  concordant_count <- 0
  discordant_count <- 0
  num_rows <- dim(x)[1]
  num_cols <- dim(x)[2]

  for (col_idx in 1:(num_cols - 1)) {
    for (row_idx in 1:(num_rows - 1)) {
      concordant_count <- concordant_count + sum(x[(row_idx + 1):num_rows, (col_idx + 1)
      discordant_count <- discordant_count + sum(x[1:row_idx, (col_idx + 1):num_cols]) *
    }
  }

  return((concordant_count - discordant_count) / (concordant_count + discordant_count))
}
```

7.3 Miara ϕ

Miara ϕ określa siłę zależności w tabelach dwuwymiarowych. Definiowana jest jako:

$$\phi = \sqrt{\frac{X^2}{n}}$$

gdzie X^2 to statystyka chi-kwadrat, a n to liczba obserwacji.

```
fi <- function(x){
  total_count <- sum(x)

  num_rows <- dim(x)[1]
  num_cols <- dim(x)[2]

  col_totals <- numeric(num_cols)
  row_totals <- numeric(num_rows)

  row_totals <- rowSums(x)
  col_totals <- colSums(x)

  X_value <- 0
```

```
  for (i in 1:num_rows){
    for (j in 1:num_cols){
      expected_value <- (row_totals[i] * col_totals[j]) / total_count
      deviation <- x[i, j] - expected_value
      X_value <- X_value + (deviation^2 / expected_value)
    }
  }
  return(sqrt(X_value/total_count))
}
```

7.4 Współczynnik Sommersa $\hat{d}$

Współczynnik Sommersa $\hat{d}$ mierzy asymetryczną zależność między zmiennymi:

$$\hat{d} = \frac{C - D}{\frac{n(n-1)}{2} - T_1}$$

gdzie $T_1 = \sum_{i=1}^{R} \frac{n_{i+}(n_{i+}-1)}{2}$.

```
sommers_d <- function(x) {
  # Obliczenie liczby wierszy i kolumn
  num_rows <- dim(x)[1]
  num_cols <- dim(x)[2]

  # Obliczenie liczby par zgodnych (C) i niezgodnych (D)
  C <- 0
  D <- 0
  for (j in 1:(num_cols - 1)) {
    for (i in 1:(num_rows - 1)) {
      C <- C + sum(x[(i + 1):num_rows, (j + 1):num_cols]) * x[i, j]
      D <- D + sum(x[1:i, (j + 1):num_cols]) * x[i + 1, j]
    }
  }

  # Obliczenie n (całkowita liczba obserwacji)
  n <- sum(x)

  # Obliczenie T_1
  row_totals <- rowSums(x)
  T1 <- sum(row_totals * (row_totals - 1) / 2)

  # Obliczenie współczynnika d_b
  db <- (C - D) / (n * (n - 1) / 2 - T1)
  return(db)
}
```

Tabela 5: Miary zależności dla różnych tabelek

Tabela	Tau	Gamma	Fi	Sommers_d
Female i WageCat	0.0477604	-0.5350192	0.3781851	-0.4204032
EducCat i WageCat	0.0426829	0.4776232	0.3571739	0.3704699
EducCat2 i WageCat	0.0694788	0.4477559	0.4558364	0.3416075

Na podstawie wyników z tabeli, obliczone miary zależności dostarczają cennych informacji o sile oraz kierunku zależności pomiędzy zmiennymi w różnych tabelach kontyngencji.

Z analizy wyników wynika, że:

- Miara γ najlepiej oddaje zależność między zmiennymi w tabelach "EducCat i WageCat" oraz "EducCat2 i WageCat", wskazując na pozytywną zależność, szczególnie w przypadku wykształcenia i kategorii dochodów.
- Wartości τ są bardzo małe we wszystkich tabelach, co sugeruje stosunkowo słabą zależność między zmiennymi. Najwyższe wartości τ są dla "EducCat2 i WageCat", wskazując na pewną pozytywną zależność.
- Miara ϕ oraz Sommersa $\hat{d}$ również wskazują na najsilniejszą zależność w przypadku "EducCat2 i WageCat", sugerując silny wpływ wykształcenia na kategorię dochodów.

Zatem w przypadku tabeli "EducCat2 i WageCat", miary takie jak ϕ i γ wskazują na najsilniejszą zależność, co może oznaczać, że wykształcenie ma bardziej wyraźny wpływ na kategorię dochodów w porównaniu do płci (w przypadku "Female i WageCat").

8 Zadanie 8

W tej sekcji przeprowadzono analizę korespondencji pomiędzy zmiennymi `wageCat` i `educCat2`. Wyniki przedstawiają się następująco:

8.1 Współrzędne punktów

W wyniku analizy korespondencji otrzymaliśmy współrzędne punktów dla zmiennych `wageCat` i `educCat2`, które przedstawiają pozycje tych zmiennych w przestrzeni 2D. Współrzędne punktów reprezentują pozycje poszczególnych kategorii zmiennych `wageCat` i `educCat2` w przestrzeni o obniżonej wymiarowości, co pozwala na wizualizację zależności między kategoriami w przestrzeni 2D.

Wiersze (`wageCat`) Współrzędne kategorii tej zmiennej ukazują, w jakim stopniu różne przedziały wynagrodzeń różnicują się pod względem edukacji.

Tabela 6: Współrzędne dla wierszy (wageCat)

Kategoria	Dim 1	Dim 2	Dim 3
1	1.3883	0.9184	-0.4292
2	0.3219	-1.6890	-0.2428
3	-0.3881	0.2781	1.6695
4	-1.3432	0.4785	-0.9910

Kolumny (`educCat2`) Współrzędne dla poziomów edukacji ukazują ich relacje względem przedziałów wynagrodzeń.

Tabela 7: Współrzędne dla kolumn (educCat2)

Kategoria	Dim 1	Dim 2	Dim 3
[0,8]	0.9980	1.0132	-3.1516
(8,11]	1.6096	1.0183	1.2699
(11,12]	0.1997	-0.4242	0.0767
(12,14]	-0.2107	-1.6587	-0.0890
(14,18]	-1.5201	0.9890	0.1880

8.2 Macierz ładunków

Macierz ładunków wskazuje, jaką wagę każda kategoria zmiennej ma w analizie korespondencji.

Masy wierszy (`wageCat`)

Wartości te wskazują, jak często poszczególne przedziały wynagrodzeń występują w danych. Większe wartości oznaczają kategorie częściej występujące, co ma większy wpływ na pozycjonowanie punktów w przestrzeni 2D.

Tabela 8: Masy dla wierszy (wageCat)

Kategoria	Masy
1	0.2529
2	0.2490
3	0.2490
4	0.2490

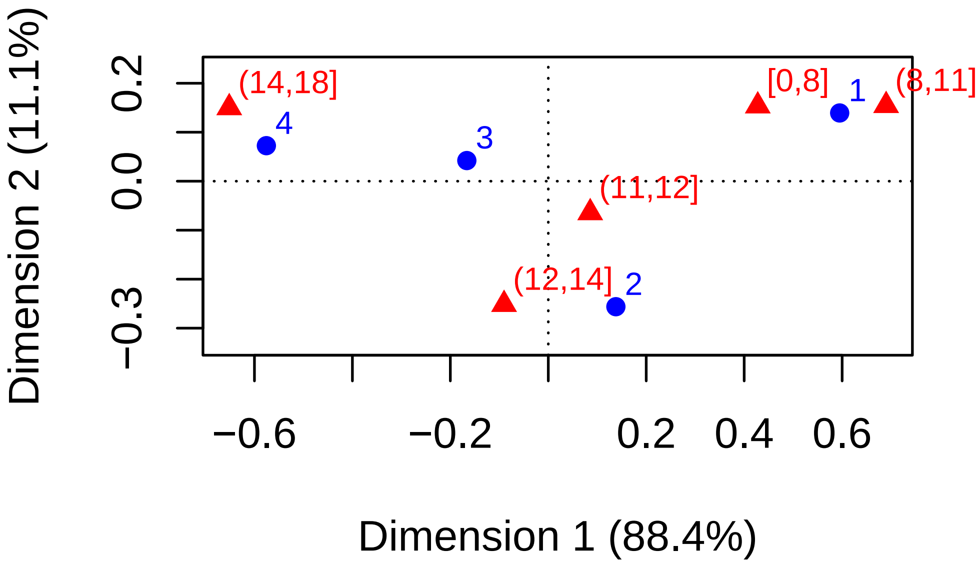
Masy kolumn (`educCat2`) Analogicznie, masy dla poziomów edukacji ukazują ich względne znaczenie w analizie.

Tabela 9: Masy dla kolumn (educCat2)

Kategoria	Masy
[0,8]	0.0760
(8,11]	0.1445
(11,12]	0.3764
(12,14]	0.1749
(14,18]	0.2281

8.3 Wykres analizy korespondencji

Aby lepiej zobaczyć zależności pomiędzy kategoriami zmiennych, generujemy wykres analizy korespondencji:



Rysunek 1: Analiza korespondencji dla wageCat i educCat2

Układ punktów na wykresie pokazuje podobieństwa i różnice pomiędzy kategoriami obu zmiennych. Kategorie leżące blisko siebie można interpretować jako mające podobne rozkłady w macierzy kontyngencji. Wykres pozwala także zauważyć, czy poszczególne kategorie zmiennych grupują się w klastery, co może świadczyć o wspólnych cechach badanych grup.

8.4 Podsumowanie

Analiza korespondencji pomiędzy zmiennymi `wageCat` i `educCat2` dostarcza cennych informacji o zależnościach pomiędzy tymi zmiennymi. Dzięki macierzy kontyngencji oraz współrzędnym punktów w przestrzeni 2D możemy lepiej zrozumieć, jak kategorie tych zmiennych się ze sobą łączą.