Sprawozdanie 2

Kacper Równicki, Wojciech Strycharczyk

22 maja 2022

1 Lista 5, 6, 7

Część pierwsza sprawozdania

1.1 Zadanie 2

W każdym z podpunktów wykonamy trzy testy - `prop.test` w dwóch wersjach (z poprawką oraz bez poprawki – jest to poprawka ciągłości Yates'a) oraz `binom.test`. Dla testów zakładamy poziom ufności 0.95.

1.1.1 a.

Niech p oznacza prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap.

W podpunkcie a, hipotezą zerową jest

$$H_0 : p \leq \frac{1}{4}.$$

Zatem hipotezą alternatywną jest

$$H_a : p > \frac{1}{4}.$$

[10]:
```
prop.test(44, 200, p = 1/4, alternative = "g")
prop.test(44, 200, p = 1/4, alternative = "g", correct = FALSE)
binom.test(44, 200, p = 1/4, alternative = "g")
```

```
1-sample proportions test with continuity correction

data:  44 out of 200, null probability 1/4
X-squared = 0.80667, df = 1, p-value = 0.8154
alternative hypothesis: true p is greater than 0.25
95 percent confidence interval:
 0.1734603 1.0000000
sample estimates:
   p
0.22
```

```
1-sample proportions test without continuity correction

data:  44 out of 200, null probability 1/4
X-squared = 0.96, df = 1, p-value = 0.8364
alternative hypothesis: true p is greater than 0.25
95 percent confidence interval:
 0.1757337 1.0000000
sample estimates:
   p
0.22



Exact binomial test

data:  44 and 200
number of successes = 44, number of trials = 200, p-value = 0.8562
alternative hypothesis: true probability of success is greater than 0.25
95 percent confidence interval:
 0.172679 1.000000
sample estimates:
probability of success
                  0.22
```

W teście `prop.test` bez korekcji p-wartość kształtuje się na poziomie $0.8154 > 0.05$, natomiast z korekcją $0.8364 > 0.05$. W `binom.test` wynosi ona $0.8562 > 0.05$, więc nie ma przesłanek do odrzucenia H_0.

1.1.2 b.

Ponownie, niech p oznacza prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap.

W podpunkcie b, hipotezą zerową jest

$$H_0 : p = \frac{1}{2}.$$

Zatem hipotezą alternatywną jest

$$H_a : p \neq \frac{1}{2}.$$

[6]:
```r
prop.test(44, 200, p = 1/2, alternative = "t")
prop.test(44, 200, p = 1/2, alternative = "t", correct = FALSE)
binom.test(44, 200, p = 1/2, alternative = "t")
```

```
1-sample proportions test with continuity correction

data:  44 out of 200, null probability 1/2
X-squared = 61.605, df = 1, p-value = 4.198e-15
```

```
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1659406 0.2850661
sample estimates:
   p
0.22
```

```
1-sample proportions test without continuity correction

data:  44 out of 200, null probability 1/2
X-squared = 62.72, df = 1, p-value = 2.383e-15
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1681654 0.2823880
sample estimates:
   p
0.22
```

```
Exact binomial test

data:  44 and 200
number of successes = 44, number of trials = 200, p-value = 6.838e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1646361 0.2838612
sample estimates:
probability of success
                  0.22
```

W tym przypadku wszystkie p-wartości są mniejsze od 0.05. Odpowiednio dla `prop.test` bez korekcji, z korekcją oraz dla `binom.test` wynoszą one $4.198 \cdot 10^{-15}$, $2.383 \cdot 10^{-15}$ oraz $6.838 \cdot 10^{-16}$. Odrzucamy więc hipotezę zerową.

1.1.3 c.

W podpunkcie c rozpatrujemy lek Ibuprom, więc niech p będzie prawdopodobieństwem, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Ibuprom.

Z zadania odczytujemy hipotezę zerową jako

$$H_0 : p \geq \frac{1}{5}$$

oraz hipotezę alternatywną jako

$$H_a : p < \frac{1}{5}.$$

[4]:
```
prop.test(35, 200, p = 1/5, alternative = "l")
prop.test(35, 200, p = 1/5, alternative = "l", correct = FALSE)
binom.test(35, 200, p = 1/5, alternative = "l")
```

```
1-sample proportions test with continuity correction

data:  35 out of 200, null probability 1/5
X-squared = 0.63281, df = 1, p-value = 0.2132
alternative hypothesis: true p is less than 0.2
95 percent confidence interval:
 0.0000000 0.2261568
sample estimates:
    p
0.175



1-sample proportions test without continuity correction

data:  35 out of 200, null probability 1/5
X-squared = 0.78125, df = 1, p-value = 0.1884
alternative hypothesis: true p is less than 0.2
95 percent confidence interval:
 0.0000000 0.2234492
sample estimates:
    p
0.175



Exact binomial test

data:  35 and 200
number of successes = 35, number of trials = 200, p-value = 0.2151
alternative hypothesis: true probability of success is less than 0.2
95 percent confidence interval:
 0.0000000 0.2252414
sample estimates:
probability of success
                 0.175
```

W przypadku p-wartości są większe od 0.05, (w kolejności: 0.2132, 0.1884, 0.2151). Nie ma więc podstaw do odrzucania hipotezy zerowej.

1.1.4 d.

W podpunkcie d powtarzamy testy z podpunktów a-c, jednak rozpatrujemy tylko badaną populację do lat 35.

	prop.test (z poprawką)	prop.test (bez poprawki)	binom.test
a)	0.5	0.55	0.59
b)	0.0	0.00	0.00
c)	1.0	1.00	1.00

Wyniki z podpunktu a) oraz c) mają p-wartość większą od 0.05, więc nie mamy podstaw do odrzucenia hipotez zerowych. W podpunkcie b) wszystkie p-wartości są mniejsze od 0.05, więc odrzucamy hipotezę zerową.

1.2 Zadanie 3

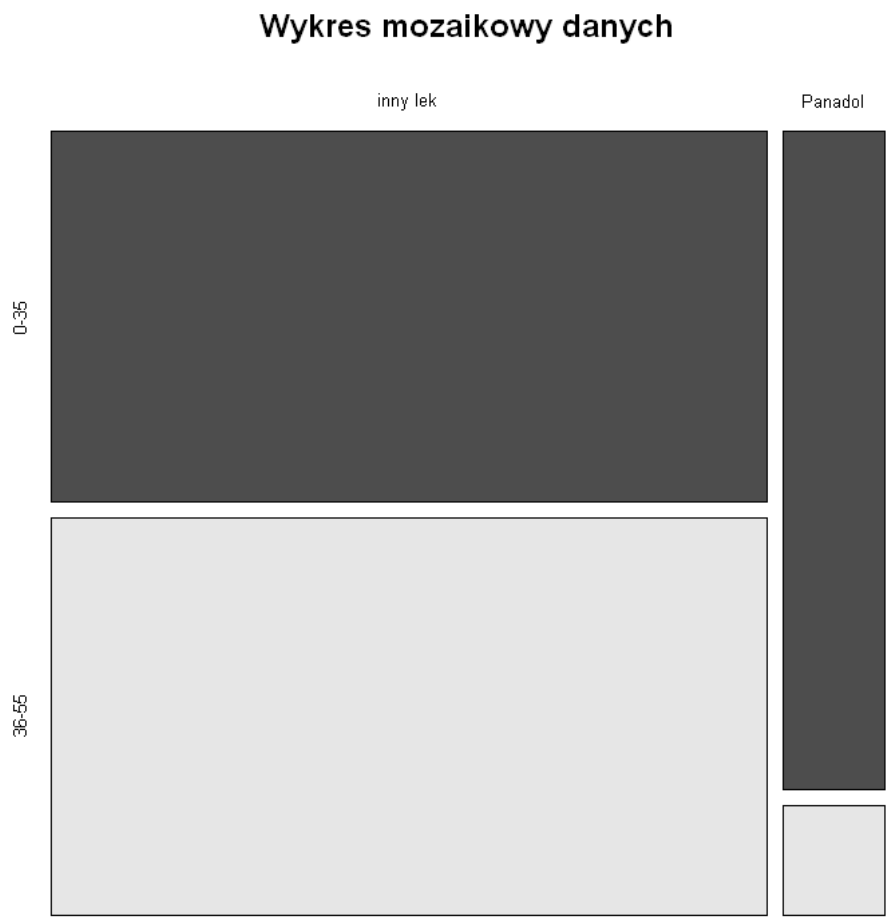
W zadaniu trzecim, musimy przede wszystkim ograniczyć tabelę z poprzedniego zadania. Rozpatrujemy tylko lek Panadol oraz grupę zbiorczą "inny lek" oraz grupy wiekowe $0-35$ i $36-55$.

Następnie przeprowadzimy test Fishera (`fisher.test`) na poziomie istotności 0.05 dla tabeli dwudzielczej przedstawionej poniżej.

Zanim jednak przeprowadzimy test, wprowadźmy hipotezę zerową i alternatywną: * H_0 : zmienne są niezależne, czyli nie ma żadnego związku pomiędzy dwiema zmiennymi kategorycznymi. Innymi słowy, znając wartość jednej zmiennej, nie możemy przewidzieć wartości drugiej zmiennej, * H_a : zmienne są zależne, czyli istnieje związek pomiędzy dwiema zmiennymi kategorycznymi. Innymi słowy, znając wartość jednej zmiennej, możemy przewidzieć wartości drugiej zmiennej.

	0-35	36-55
inny lek	72	77
Panadol	18	3

Zależność między tymi zmiennymi łatwo zobrazować za pomocą wykresu mozaikowego.



Widzimy, że możwmy spodziewać się zależności. Przeprowadźmy zatem test Fishera, który rozwieje nasze wątpliwości.

[8]:
```
fisher.test(data)
```

```
Fisher's Exact Test for Count Data

data:  data
p-value = 0.001789
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.0285040 0.5718021
sample estimates:
odds ratio
```

```
0.1573343
```

Zgodnie z oczekiwaniami, uzyskaliśmy p-wartość na poziomie $0.002 < 0.05$, więc odrzucamy hipotezę zerową. Zmienne zatem są **zależne**.

Zarówno w tym teście, jak i w testach z zadania poprzedniego, hipotezy są takie same. Na podstawie wyników tamtego testu można więc wyciągnąć analogiczny wniosek jak w teście Fishera.

1.3 Zadanie 4

W zadaniu czwartym weryfikujemy hipotezę o niezależności stopnia zadowolenia z pracy i wynagrodzenia na podstawie danych w tablicy zamieszczonej poniżej.

Hipotezą zerową H_0 jest, że w.w. zmienne są niezależne, a hipoteza alternatywna H_a stanowi, że są zależne. Wywołamy test chi-kwadrat (`chisq.test`) na poziomie ufności 0.95, aby zobaczyć, czy odrzucamy H_0.

	b. niezadow.	niezadow.	zadow.	b. zadow.
do 6000	20	24	80	82
6000-15000	22	38	104	125
15000-20000	13	28	81	113
powyżej 25000	7	18	54	92

[10]:
```
chisq.test(data)
```

```
Pearson's Chi-squared test

data:  data
X-squared = 11.989, df = 9, p-value = 0.214
```

P-wartość wynosi $0.214 > 0.05$, więc nie mamy podstaw, żeby odrzucić hipotezę zerową. Zatem możemy powiedzieć, że zmienne są niezależne.

1.4 Zadanie 5

Hipoteza zerowa mówi nam, że zmienne są niezależne, czyli

$$H_0 : p \in \mathcal{P}_0,$$

a hipoteza alternatywna stanowi o ich zależności, więc

$$H_a : p \in \mathcal{P}_1 = \mathcal{P} \backslash \mathcal{P}_0,$$

gdzie $P_o \subset \mathcal{P} = \{p : p_i \geq 0, i = 1, \ldots, k, \sum_{i=1}^{k} p_i = 1\}$.

Poniżej prezentujemy funckję, która dla danyh w tablicy dwudzielczej oblicza wartość poziomu krytycznego (p-value) w asymptotycznym teście niezależności opartym na ilorazie wiarogodności.

[11]:
```r
p_val <- function(table) {
    n <- sum(table)

    R <- dim(table)[1]
    C <- dim(table)[2]

    Gsq <- 1

    for (i in 1:R) {
        for (j in 1:C) { #ni+ * n+j
            Gsq <- Gsq * ((sum(table[i, ]) * sum(table[, j])) / (n*table[i,␣
 ↪j]))^table[i, j]
        }
    }


    Gsq <- -2*log(Gsq)

    p_val <- 1 - pchisq(Gsq, (R-1)*(C-1))
    p_val
}
```

Wywołując tę funkcję na danych z poprzedniego zadania, sprawdźmy jej działanie.

[12]:
```r
p_val(data) %>% round(3)
```

0.211

P-wartość wynosi 0.21, więc tyle samo co w zadaniu czwartym. Nasza funkcja działa więc poprawnie.

---

2 Lista 8, 9

2.1 Zadanie 1

W zadaniu pierwszym zaimplementujemy poznane na wykładzie miary współzmienności.

Zacznijmy jednak od wprowadzenia danych z wykładu. Tak wyglądają odpowiednio dane dotyczące *Segregacja* i *Wiek* oraz *Segregacja* i *Miejsce zamieszkania*.

	A	B	C	D
18-25	888	369	50	457
26-35	263	95	10	99
36-45	208	29	2	44
46-59	78	9	0	19
60+	1	0	0	4

	Wieś	Miasto do 20 tys.	Miasto 20-50 tys.	Miasto pow. 50 tys.
A	505	202	19	136
B	240	77	14	88
C	181	63	8	105
D	512	159	21	294

W pierwszej kolejności zajmiemy się analizą zależności wieku respondentów z udzielonymi odpowiedziami. Zweryfikujmy hipotezę o niezależności zmiennych. Nasza hipoteza ma postać

$$H_0 : p_{ij} = p_{i+}p_{+j} \forall i \in \{1, \ldots, R\}, \forall j \in \{1, \ldots, C\}$$

Przeciwko hipotezie alternatywnej:

$$H_1 : p_{ij} \neq p_{i+}p_{+j}$$

Dla conajmniej jednej pary i, j, gdzie $i \in \{1, \ldots, R\}$ oraz $j \in \{1, \ldots, C\}$

Weryfikacja hipotezy o niezależności jest równoważna weryfikacji hipotezy o jednorodności rozkładów warunkowych, tzn. równości rozkładów. Przeprowadzimy test na poziomie istotności $\alpha = 0.05$.

[581]:
```
fisher.test(data_wiek, simulate.p.value = T)$p.value
```

0.000499750124937531

[582]:
```
chisq.test(data_wiek)$p.value
```

```
Warning message in chisq.test(data_wiek):
''Aproksymacja chi-kwadrat może być niepoprawna''
```

9.14122959499407e-13

Jak widać wartość krytyczna w obu testach jest znacznie mniejsza od przyjętego poziomu $\alpha = 0.05$, zatem odrzucamy hipotezę o niezależności zmiennych.

Teraz możemy zaimplementować miary. Dla każdej z nich, napisaliśmy oddzielną funkcję, która jest wywoływana przez funcję `coefficient`.

[583]:
```
#1
goodman <- function(data) {
  n <- sum(data)

  R <- dim(data)[1]
  C <- dim(data)[2]

  sum1 <- 0

  for (i in 1:R) {
    for (j in 1:C) {
      sum1 <- sum1 + data[i, j]^2 / (n * sum(data[i, ]))
    }
  }
```

```
  sum2 <- 0
  for (j in 1:C) {
    sum2 <- sum2 + (sum(data[ , j])/n)^2
  }

  (sum1 - sum2) / (1 - sum2)
}

#2
crammer <- function(data) {
  chi2 <- as.numeric(chisq.test(data, simulate.p.value = TRUE)$statistic)
  C <- dim(data)[2]
  R <- dim(data)[1]
  sqrt(chi2/(sum(data)*min(R-1, C-1)))

}

#3
t.czuprow <- function(data) {
  chi2 <- as.numeric(chisq.test(data, simulate.p.value = TRUE)$statistic)
  C <- dim(data)[2]
  R <- dim(data)[1]
  sqrt(chi2/(sum(data)*sqrt((R-1)*(C-1))))
}

#4
phi <- function(data) {
  chi2 <- as.numeric(chisq.test(data, simulate.p.value = TRUE)$statistic)
  sqrt(chi2/sum(data))
}

# 5
pearson <- function(data) {
  chi2 <- as.numeric(chisq.test(data, simulate.p.value = TRUE)$statistic)
  sqrt(chi2/(chi2 + sum(data)))

}

coefficient <- function(data, coef) {
  if (coef == 1) {
    goodman(data)
  } else if (coef == 2) {
    crammer(data)
  } else if (coef == 3) {
    t.czuprow(data)
  } else if (coef == 4) {
    phi(data)
```

```
  } else if (coef == 5) {
    pearson(data)
  }
}
```

Sprawdźmy zatem, czy nasze funkcje działą poprawnie. Porównamy ich wyniki z wynikami obliczonymi przez funkcje wbudowane w pakiecie R.

[584]:
```
coefficient(data_wiek, 1)
GoodmanKruskalTau(data_wiek, direction = "column")
```

0.0171216260083961

0.0171216260083962

[585]:
```
coefficient(data_wiek, 2)
CramerV(data_wiek)
```

0.10292408364219

0.10292408364219

[586]:
```
coefficient(data_wiek, 3)
TschuprowT(data_wiek)
```

0.0957816523560532

0.0957816523560532

[587]:
```
coefficient(data_wiek, 4)
phi(data_wiek)
```

0.178269742190742

0.178269742190742

[588]:
```
coefficient(data_wiek, 5)
ContCoef(data_wiek)
```

0.175502805201971

0.175502805201971

Jak widać, w każdym przypadku nasza funkcja zwraca dokładnie taką samą wartość.

2.1.1 Analiza korespondencji

Zdefinujmy funkcje, które będą obliczać wektory: *przeciętny profil wierszowy*, *przeciętny profil kolumnowy* oraz macierze: *macierz profili wierszowych*, *macierz profili kolumnowych*.

[589]:
```
count_average_row_profile <- function(data) {
  rows <- numeric(nrow(data))
  for (i in 1:nrow(data)) {
    rows[i] <- sum(data[i, ])
```

```
  }
  rows
}
count_average_col_profile <- function(data) {
  cols <- numeric(ncol(data))
  for (i in 1:ncol(data)) {
    cols[i] <- sum(data[, i])
  }
  cols
}
count_row_massess_matrix <- function(data) {
    result <- matrix(0, nrow(data), ncol(data))
  for (i in 1:nrow(data)) {
      for (j in 1:ncol(data)) {
            result[i, j] <-data[i,j]/sum(data[i, ])
      }
  }
  result
}
count_col_massess_matrix <- function(data) {
    result <- matrix(0, nrow(data), ncol(data))
      for (i in 1:nrow(data)) {
          for (j in 1:ncol(data)) {
                result[i, j] <- data[i,j]/sum(data[, j])
          }
      }
    result
}
```

Następnie wyznaczamy *macierz korespondencji* **P** (macierz częstości zaobserwawanych) oraz przeciętne profile.

[590]:
```
n <- sum(data_wiek)
P <- data_wiek/n

r <- count_average_row_profile(P)
c <- count_average_col_profile(P)
```

Macierz częstości wierszowych wygląda następująco:

$$D_r = diag(r)$$

[591]:
```
Dr <- diag(length(r)) * r
Dr
```

0.672	0.0000000	0.0000000	0.00000000	0.000000000
0.000	0.1779048	0.0000000	0.00000000	0.000000000
0.000	0.0000000	0.1078095	0.00000000	0.000000000
0.000	0.0000000	0.0000000	0.04038095	0.000000000
0.000	0.0000000	0.0000000	0.00000000	0.001904762

Natomiast macierz częstości kolumnowych:

$$D_c = diag(c)$$

[592]:
```
Dc <- diag(length(c)) * c
Dc
```

0.5478095	0.0000000	0.00000000	0.0000000
0.0000000	0.1912381	0.00000000	0.0000000
0.0000000	0.0000000	0.02361905	0.0000000
0.0000000	0.0000000	0.00000000	0.2373333

Macierz profili wierszowych:

[593]:
```
R = count_row_massess_matrix(data_wiek)
R
```

0.5034014	0.20918367	0.028344671	0.2590703
0.5631692	0.20342612	0.021413276	0.2119914
0.7349823	0.10247350	0.007067138	0.1554770
0.7358491	0.08490566	0.000000000	0.1792453
0.2000000	0.00000000	0.000000000	0.8000000

Macierz profili kolumnowych:

[594]:
```
C <- count_col_massess_matrix(data_wiek)
C
```

0.6175243394	0.73505976	0.80645161	0.733547352
0.1828929068	0.18924303	0.16129032	0.158908507
0.1446453408	0.05776892	0.03225806	0.070626003
0.0542420028	0.01792829	0.00000000	0.030497592
0.0006954103	0.00000000	0.00000000	0.006420546

Następnie obliczyliśmy macierz rezyduów standaryzowanych zgodnie ze wzorem:

$$A = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$$

[595]:
```
A <- solve(Dr ^ (1/2)) %*% (P - r %*% t(c)) %*% solve(Dc ^ (1/2))
A
```

-0.049184998	0.03363991	0.025206501	0.03657666
0.008753063	0.01175547	-0.006053731	-0.02194086
0.083034109	-0.06664708	-0.035362705	-0.05516979
0.051053163	-0.04886150	-0.030882999	-0.02396049
-0.020509120	-0.01908568	-0.006707359	0.05040715

W analizie korespondencji zastosowaliśmy podejście Greenacre'a, które polega na dekompozycji według wartości osobliwych macierzy **A**. W tym celu wykorzystaliśmy wbudowaną funckję *svd*, aby otrzymać macierze U, Γ, V, zgdonie ze wzorem:

$$A = U\Gamma V^T$$

Po wyznaczeniu rozkładu macierzy **A** według wartości osobliwych, możemy wyznaczyć macierze **F** i **G**, nazywamy współrzędnymi kategorii cech odpowiednio dla wierszy i kolumn.

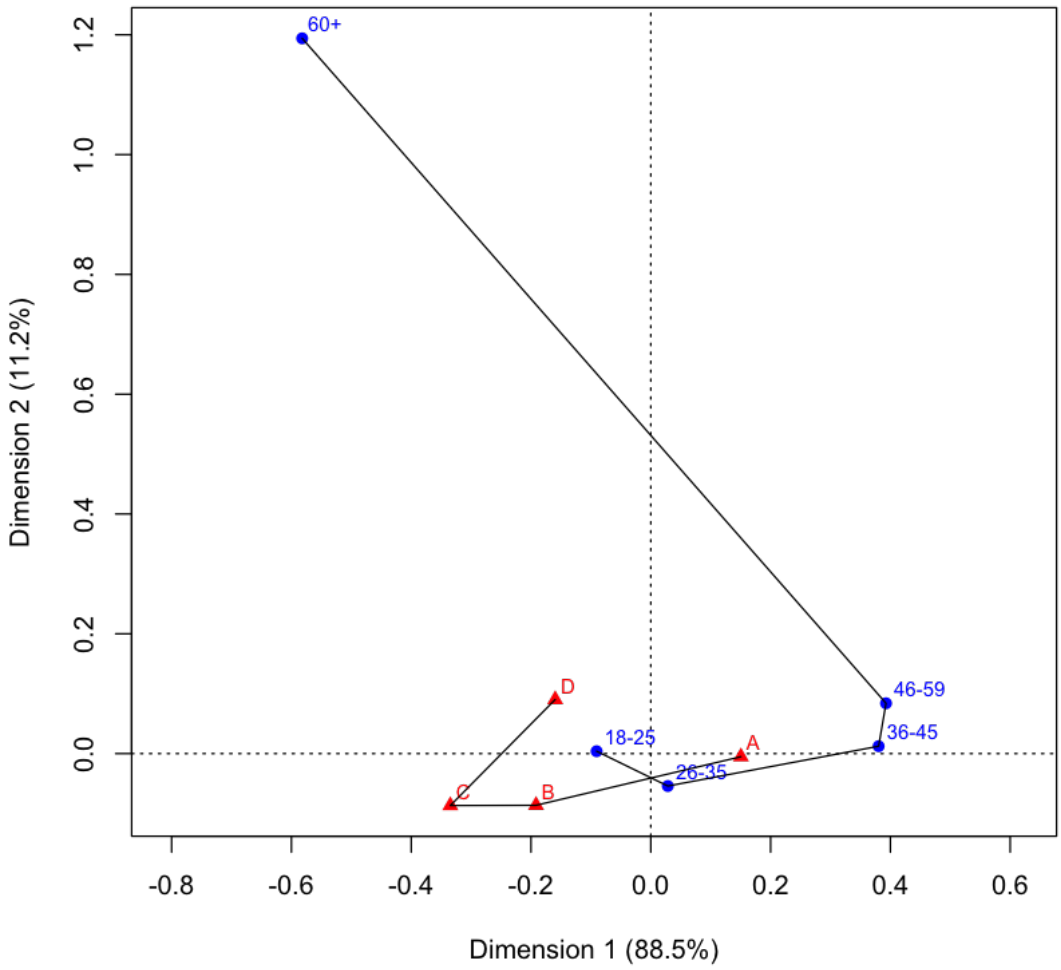
[596]:
```
sv <- svd(A)
gamm <- diag(length(sv$d)) * sv$d
U <- sv$u
V <- sv$v

F <- solve(Dr ^ (1/2)) %*% U %*% gamm
G <- solve(Dc ^ (1/2)) %*% V %*% gamm

F
G
```

-0.09048091	0.003935358	-0.00424245	-6.327408e-17
0.02838125	-0.054172767	0.01947277	-4.863980e-17
0.38035147	0.012314854	-0.01287347	-3.971103e-17
0.39268765	0.083974027	0.01581285	-1.196198e-16
-0.58201684	1.194071935	0.07138559	6.898849e-17

0.1504620	-0.005221131	-0.001210770	6.219487e-17
-0.1915971	-0.086311190	0.009831168	6.219487e-17
-0.3347124	-0.086721499	-0.064268200	6.219487e-17
-0.1595994	0.090229432	0.001268811	6.219487e-17

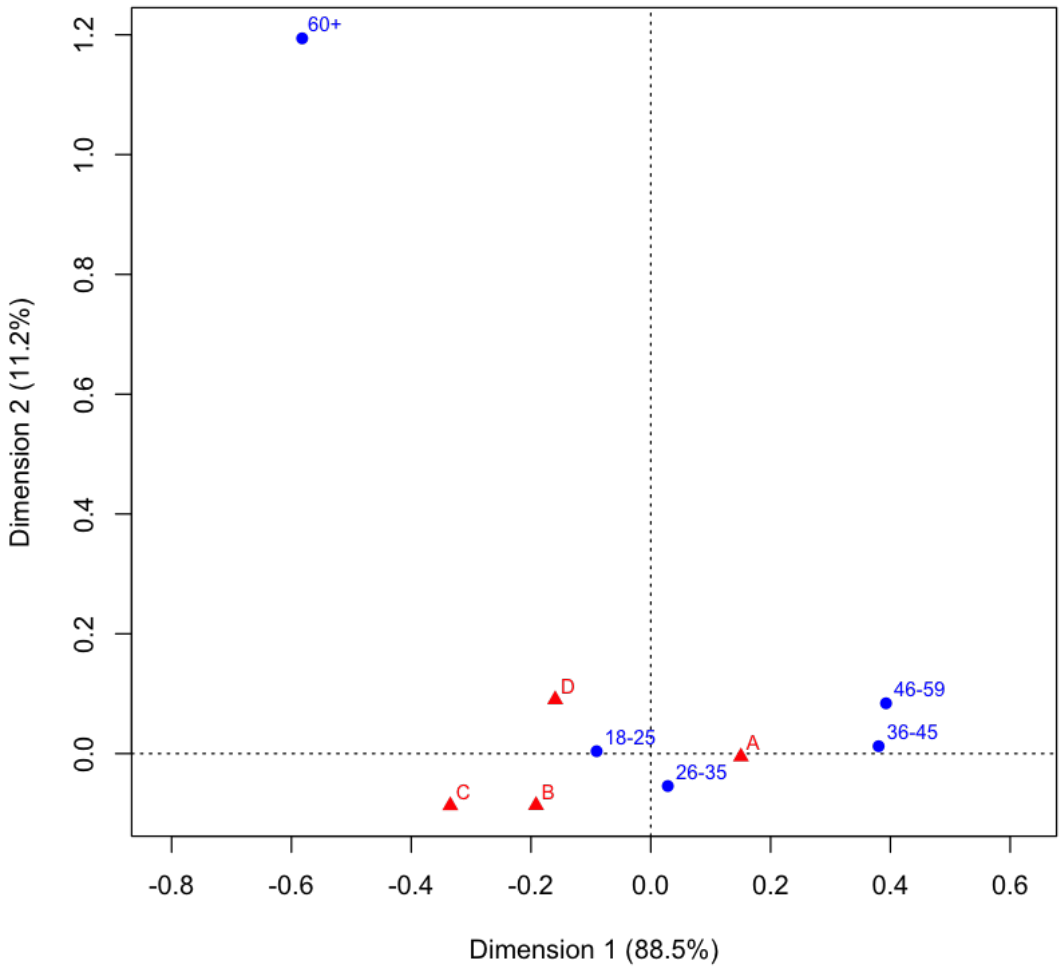
Aby sprawdzić poprawność naszej metody, porównamy wykres wygenerowany przez wbudowaną funkcję *ca* z wykresem stworzonym przez dane pochodzące z obliczonych macierzy F i G. Aby zachować przjrzystość na wykresie, punkty wygenerowane przez nas zwizulizujemy za pomocą wykresu liniowego. Wówczas punkty nie będą się pokrywać i będziemy mogli zweryfikować poprawność naszej metody.

[597]:
```
plot(ca(data_wiek))
xs <- F[,1]
ys <- F[,2]
lines(xs, ys)
xs <- G[,1]
ys <- G[,2]
lines(xs, ys)
```



Jak widać na wykresie, punkty stworzone przez funkcję *ca* są połączone liniami, zatem nasza metoda została zaimplementowana poprawnie. Sam wykres, który otrzymujemy przez funkcję *ca*, wygląda następująco:

[723]:
```
plot(ca(data_wiek))
```



2.1.2 Wnioski

Osoby w wieku 26-35 lat najczęściej segregują śmieci, ponieważ jest to korzystne dla środowiska (odpowiedź A). Na podstawie odpowiedzi osób w wieku 36-45 oraz 46-59 możemy wnioskować, że segregują śmieci również z tego samego powodu, jednak ta zależność jest mniejsza. Z wykresu możemy również odczytać, że osoby w wieku 18-25 lat w dużej części segregują śmieci ponieważ taki jest wymóg ustawowy (odpowiedź B) lub nie segregują śmieci (odpowiedź D). Udział osób powyżej 60 lat biorących udział w badaniu jest bardzo mały, stąd trudno tu znaleźć zależności.

W analizowanych danych inercja całkowita wynosi $\lambda \approx 0.032$, zatem punkty profili są rozproszone w niewielkim stopniu i skupiają się na niewielkim obszarze.

Następenie zajmiemy się danymi *Segregacja* i *Miejsce zamieszkania*, dla których obliczmy odpowiednie miary współzmienności, a także porównamy je z wynikami otrzymanymi z funkcji

wbudowanych.

[724]:
```
coefficient(data_miejsce, 1)
GoodmanKruskalTau(data_miejsce, direction = "column")
```

0.0101279237876796

0.0101279237876795

[725]:
```
coefficient(data_miejsce, 2)
CramerV(data_miejsce)
```

0.0911602380776021

0.0911602380776021

[726]:
```
coefficient(data_miejsce, 3)
TschuprowT(data_miejsce)
```

0.0911602380776021

0.0911602380776021

[727]:
```
coefficient(data_miejsce, 4)
phi(data_miejsce)
```

0.157894163980482

0.157894163980482

[728]:
```
coefficient(data_miejsce, 5)
ContCoef(data_miejsce)
```

0.155962021385397

0.155962021385397

W tym przypadku wszystkie obliczone współczynniki również zgadzają się w wynikami obliczonymi za pomocą funkcji wbudowanych.

2.2 Zadanie 2

W zadaniu drugim oraz trzecim obliczmy odpowiednie miary współzmienności oraz przeprowadzimy analizę korespondencji. Zanim jednak do niej przejdziemy, wykonamy testy: Fishera oraz chi-kwadrat Pearsona, które zweryfikują hipotezę o niezależności.

Wprowadźmy dane i zobaczmy jak wyglądają dane wejściowe:

	do lat 35	od 36 do 55	powyzej 55
Ibuprom	35	0	0
Apap	22	22	0
Paracetamol	15	15	15
Ibuprofen	0	40	10
Panadol	18	3	5

Analogicznie jak w poprzednim zadaniu, mamy do czynienia z hipotezą:

$$H_0 : p_{ij} = p_{i+}p_{+j} \forall i \in \{1, \ldots, R\}, \forall j \in \{1, \ldots, C\}$$

Przeciwko hipotezie alternatywnej:

$$H_1 : p_{ij} \neq p_{i+}p_{+j}$$

Dla conajmniej jednej pary i, j, gdzie $i \in \{1, \ldots, R\}$ oraz $j \in \{1, \ldots, C\}$

Weryfikacja hipotezy o niezależności jest równoważna weryfikacji hipotezy o jednorodności rozkładów warunkowych, tzn. równości rozkładów. Przeprowadzimy test na poziomie istotności $\alpha = 0.05$.

[730]:
```
chisq.test(data)
```

```
Warning message in chisq.test(data):
''Aproksymacja chi-kwadrat może być niepoprawna''
```

```
Pearson's Chi-squared test

data:  data
X-squared = 114.97, df = 8, p-value < 2.2e-16
```

[731]:
```
fisher.test(data, simulate.p.value = T)
```

```
Fisher's Exact Test for Count Data with simulated p-value (based on
2000 replicates)

data:  data
p-value = 0.0004998
alternative hypothesis: two.sided
```

W tym przypadku również obie wartości krytyczne są znacznie mniejsze od przyjętego poziomu istotności $\alpha = 0.05$, zatem odrzucamy hipotezę o niezależności zmiennych.

Następnie obliczmy miary współzmienności:

[732]:
```
coefficient(data, 1)
GoodmanKruskalTau(data, direction = "column")
```

0.347717323327079

0.347717323327079

[733]:
```
coefficient(data, 2)
CramerV(data)
```

0.536115539594658

0.536115539594658

[734]:
```
coefficient(data, 3)
TschuprowT(data)
```

0.450817635406959

0.450817635406959

[735]:
```
coefficient(data, 4)
phi(data)
```

0.758181867093736

0.758181867093736

[736]:
```
coefficient(data, 5)
ContCoef(data)
```

0.604164510101432

0.604164510101432

W tym przypadku nasza funkcja również zwraca dokładnie takie same wartości, co potwierdza poprawność implementacji poszczególnych współczynników.

2.2.1 Analiza korespondencji

Wyznaczamy *macierz korespondencji* **P** (macierz częstości zaobserwawanych) oraz przeciętne profile.

[737]:
```
n <- sum(data)
P <- data/n

r <- count_row_massess(P)
c <- count_col_massess(P)
```

Macierz częstości wierszowych wygląda następująco:

$$D_r = diag(r)$$

[626]:
```
Dr <- diag(length(r)) * r
Dr
```

0.175	0.00	0.000	0.00	0.00
0.000	0.22	0.000	0.00	0.00
0.000	0.00	0.225	0.00	0.00
0.000	0.00	0.000	0.25	0.00
0.000	0.00	0.000	0.00	0.13

Natomiast macierz częstości kolumnowych:

$$D_c = diag(c)$$

[627]:
```
Dc <- diag(length(c)) * c
Dc
```

0.45	0.0	0.00
0.00	0.4	0.00
0.00	0.0	0.15

Macierz profili wierszowych:

[628]:
```
R = count_row_massess_matrix(data)
R
```

1.0000000	0.0000000	0.0000000
0.5000000	0.5000000	0.0000000
0.3333333	0.3333333	0.3333333
0.0000000	0.8000000	0.2000000
0.6923077	0.1153846	0.1923077

Macierz profili kolumnowych:

[629]:
```
C <- count_col_massess_matrix(data)
C
```

0.3888889	0.0000	0.0000000
0.2444444	0.2750	0.0000000
0.1666667	0.1875	0.5000000
0.0000000	0.5000	0.3333333
0.2000000	0.0375	0.1666667

Następnie obliczyliśmy macierz rezyduów standaryzowanych zgodnie ze wzorem:

$$A = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$$

[630]:
```
A <- solve(Dr ^ (1/2)) %*% (P - r %*% t(c)) %*% solve(Dc ^ (1/2))
A
```

0.34298526	-0.26457513	-0.16201852
0.03496029	0.07416198	-0.18165902
-0.08249579	-0.05000000	0.22453656
-0.33541020	0.31622777	0.06454972
0.13023647	-0.16225573	0.03938632

W analizie korespondencji zastosowaliśmy podejście Greenacre'a, które polega na dekompozycji według wartości osobliwych macierzy **A**. W tym celu wykorzystaliśmy wbudowaną funckję *svd*, aby otrzymać macierze U, Γ, V, zgdonie ze wzorem:

$$A = U\Gamma V^T$$

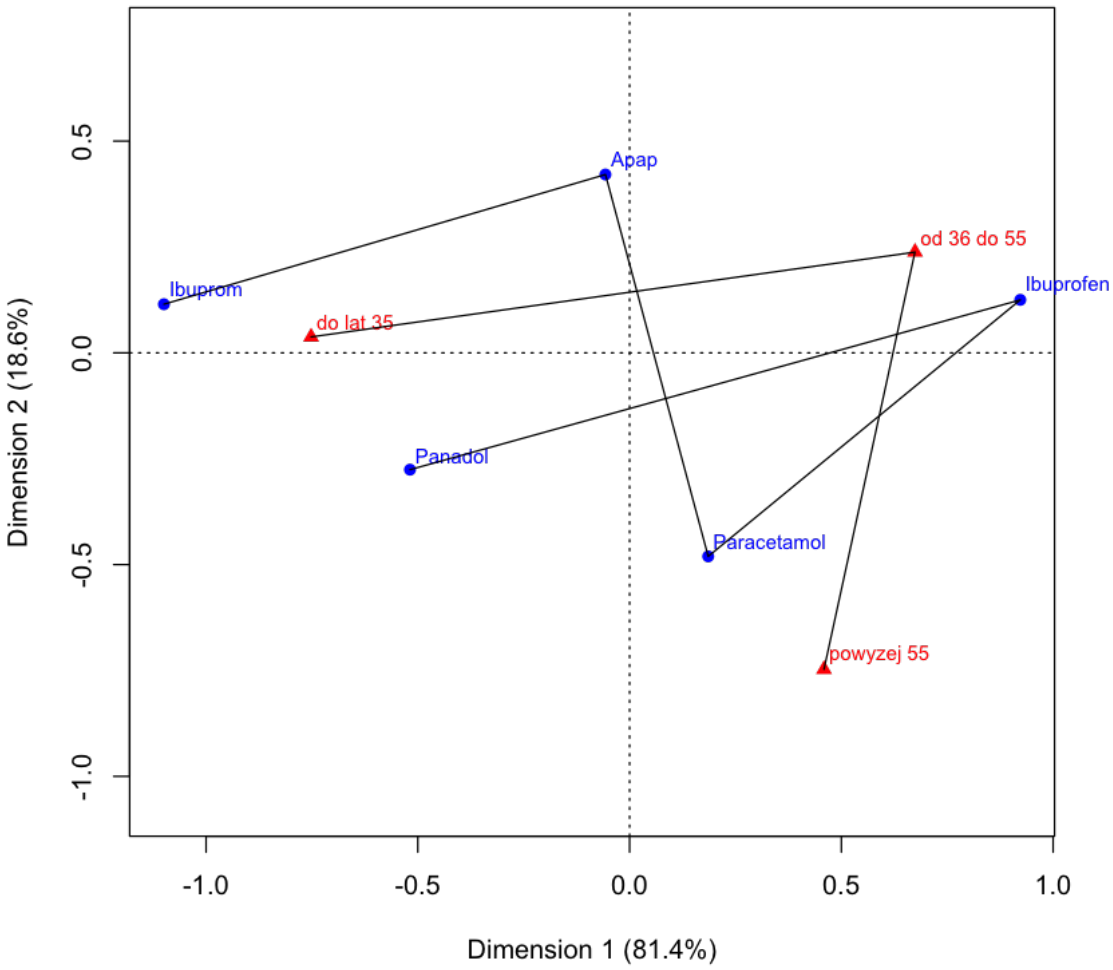
Po wyznaczeniu rozkładu macierzy **A** według wartości osobliwych, możemy wyznaczyć macierze **F** i **G**, nazywamy współrzędnymi kategorii cech odpowiednio dla wierszy i kolumn.

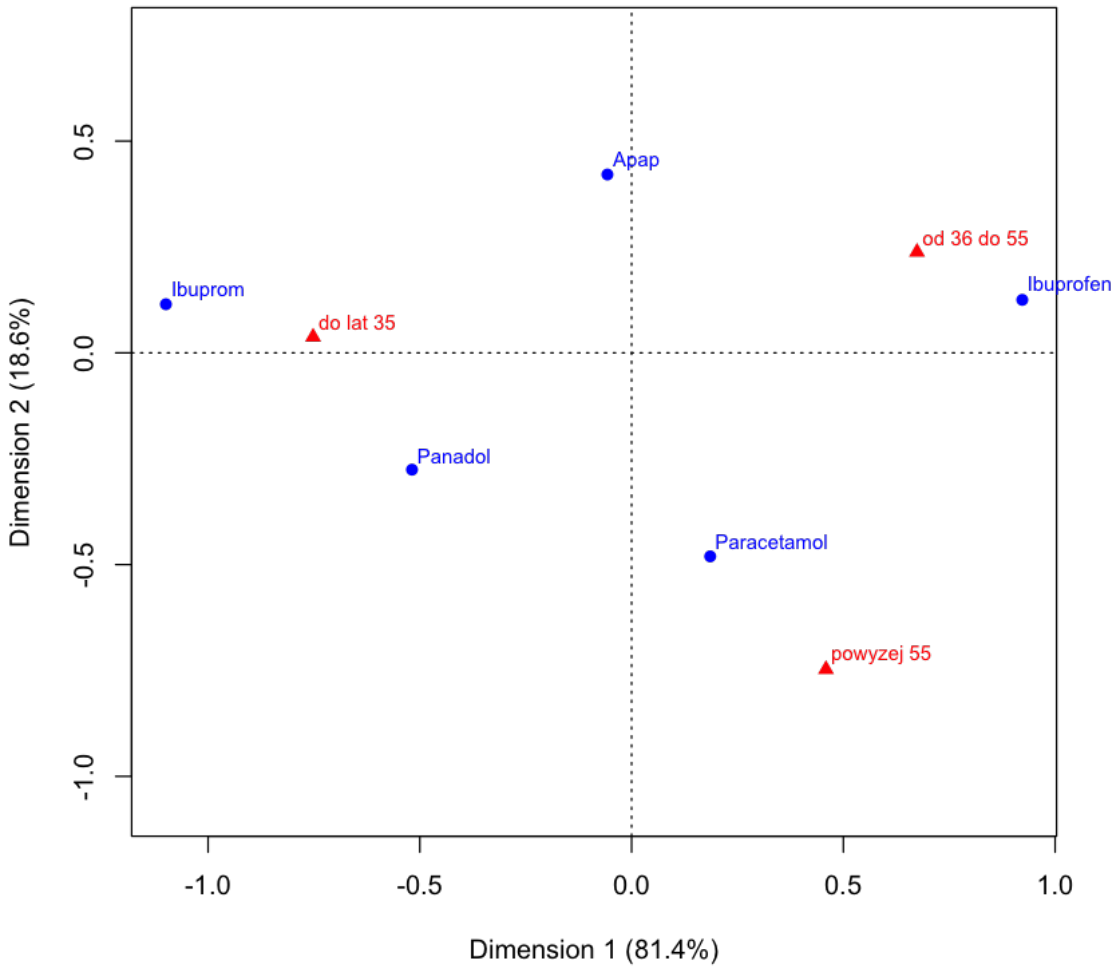
[631]:
```
sv <- svd(A)
gamm <- diag(length(sv$d)) * sv$d
U <- sv$u
V <- sv$v

F <- solve(Dr ^ (1/2)) %*% U %*% gamm
G <- solve(Dc ^ (1/2)) %*% V %*% gamm

F
G
```

-1.09956877	0.1147638	-1.717137e-16
-0.05710257	0.4210640	-4.552222e-18
0.18562161	-0.4806004	-7.665865e-17
0.92250490	0.1251055	-1.256476e-16
-0.51849220	-0.2758386	1.364674e-17

-0.7520247	0.03755489	-1.022634e-16
0.6739168	0.23802000	-1.022634e-16
0.4589629	-0.74738468	-1.022634e-16

Aby sprawdzić poprawność naszej metody, porównamy wykres wygenerowany przez wbudowaną funkcję *ca* z wykresem stworzonym przez dane pochodzące z obliczonych macierzy F i G. Aby zachować przjrzystość na wykresie, punkty wygenerowane przez nas zwizulizujemy za pomocą wykresu liniowego. Wówczas punkty nie będą się pokrywać i będziemy mogli zweryfikować poprawność naszej metody.


Apap
od 36 do 55
Ibuprom
Ibuprofen
do lat 35
Panadol
Paracetamol
powyzej 55
0.5
0.0
-0.5
-1.0
Dimension 2 (18.6%)
-1.0
-0.5
0.0
0.5
1.0
Dimension 1 (81.4%)



2.2.2 Wnioski

Osoby w wieku od 36 do 55 lat najczęściej stosują lek Ibuprofen. Możemy również podejrzewać, że duża część osób powyżej 55 roku życia stosuje Paracetamol, jednak udział klientów w tym wieku uczestniczący w badaniu jest niewielki, zatem jak widać na wykresie poszczególne punkty nie są blisko siebie. Poza tym nie widać silnych zależności.

2.3 Zadanie 3

Dane wejściowe wyglądają nastpująco:

	b. niezadow.	niezadow.	zadow.	b. zadow.
< 6000	32	44	60	70
6000 - 15000	22	38	104	125
15000 - 25000	13	48	61	113
> 25000	3	18	54	96

W pierwszej kolejności zajmiemy się analizą zależności zmiennych *Wynagrodzenie* i *Stopień zadowolenia z pracy*. Zweryfikujmy hipotezę o niezależności zmiennych. Nasza hipoteza ma postać

$$H_0 : p_{ij} = p_{i+}p_{+j} \forall i \in \{1, \ldots, R\}, \forall j \in \{1, \ldots, C\}$$

Przeciwko hipotezie alternatywnej:

$$H_1 : p_{ij} \neq p_{i+}p_{+j}$$

Dla conajmniej jednej pary i, j, gdzie $i \in \{1, \ldots, R\}$ oraz $j \in \{1, \ldots, C\}$

Weryfikacja hipotezy o niezależności jest równoważna weryfikacji hipotezy o jednorodności rozkładów warunkowych, tzn. równości rozkładów. Przeprowadzimy test na poziomie istotności $\alpha = 0.05$.

[740]:
```
fisher.test(data, simulate.p.value = T)$p.value
```

0.000499750124937531

[741]:
```
chisq.test(data)$p.value
```

4.86783132046709e-08

Wartość krytyczna w obu testach jest znacznie mniejsza od przyjętego poziomu $\alpha = 0.05$, zatem również odrzucamy hipotezę o niezależności zmiennych.

Następnie obliczmy miary współzmienności:

[742]:
```
coefficient(data, 1)
GoodmanKruskalTau(data, direction = "column")
```

0.0167118220312561

0.0167118220312562

[743]:
```
coefficient(data, 2)
CramerV(data)
```

0.13847359915759

0.13847359915759

[744]:
```
coefficient(data, 3)
TschuprowT(data)
```

0.13847359915759

0.13847359915759

[745]:
```
coefficient(data, 4)
phi(data)
```

0.239843309247872

0.239843309247872

[746]:
```
coefficient(data, 5)
ContCoef(data)
```

0.233228878890807

0.233228878890807

W tym przypadku nasza funkcja również zwraca dokładnie takie same wartości, co potwierdza poprawność implementacji poszczególnych współczynników.

2.3.1 Analiza korespondencji

Wyznaczamy *macierz korespondencji* **P** (macierz częstości zaobserwawanych) oraz przeciętne profile.

[747]:
```
n <- sum(data)
P <- data/n

r <- count_row_massess(P)
c <- count_col_massess(P)
```

Macierz częstości wierszowych wygląda następująco:

$$D_r = diag(r)$$

[748]:
```
Dr <- diag(length(r)) * r
Dr
```

0.2286349	0.0000000	0.0000000	0.0000000
0.0000000	0.3207547	0.0000000	0.0000000
0.0000000	0.0000000	0.2608213	0.0000000
0.0000000	0.0000000	0.0000000	0.1897891

Natomiast macierz częstości kolumnowych:

$$D_c = diag(c)$$

[749]:
```
Dc <- diag(length(c)) * c
Dc
```

0.07769145	0.0000000	0.0000000	0.0000000
0.00000000	0.1642619	0.0000000	0.0000000
0.00000000	0.0000000	0.3096559	0.0000000
0.00000000	0.0000000	0.0000000	0.4483907

Macierz profili wierszowych:

[750]:
```
R = count_row_massess_matrix(data)
R
```

0.15533981	0.2135922	0.2912621	0.3398058
0.07612457	0.1314879	0.3598616	0.4325260
0.05531915	0.2042553	0.2595745	0.4808511
0.01754386	0.1052632	0.3157895	0.5614035

Macierz profili kolumnowych:

[751]:
```
C <- count_col_massess_matrix(data)
C
```

0.45714286	0.2972973	0.2150538	0.1732673
0.31428571	0.2567568	0.3727599	0.3094059
0.18571429	0.3243243	0.2186380	0.2797030
0.04285714	0.1216216	0.1935484	0.2376238

Następnie obliczyliśmy macierz rezyduów standaryzowanych zgodnie ze wzorem:

$$A = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$$

[752]:
```
A <- solve(Dr ^ (1/2)) %*% (P - r %*% t(c)) %*% solve(Dc ^ (1/2))
A
```

0.133203867	0.05819913	-0.015805310	-0.07753757
-0.003183739	-0.04579816	0.051097476	-0.01341809
-0.040991673	0.05039544	-0.045963056	0.02475694
-0.094008538	-0.06341766	0.004801831	0.07352502

W analizie korespondencji zastosowaliśmy podejście Greenacre'a, które polega na dekompozycji według wartości osobliwych macierzy **A**. W tym celu wykorzystaliśmy wbudowaną funckję *svd*, aby otrzymać macierze U, Γ, V, zgdonie ze wzorem:

$$A = U\Gamma V^T$$

Po wyznaczeniu rozkładu macierzy **A** według wartości osobliwych, możemy wyznaczyć macierze **F** i **G**, nazywamy współrzędnymi kategorii cech odpowiednio dla wierszy i kolumn.

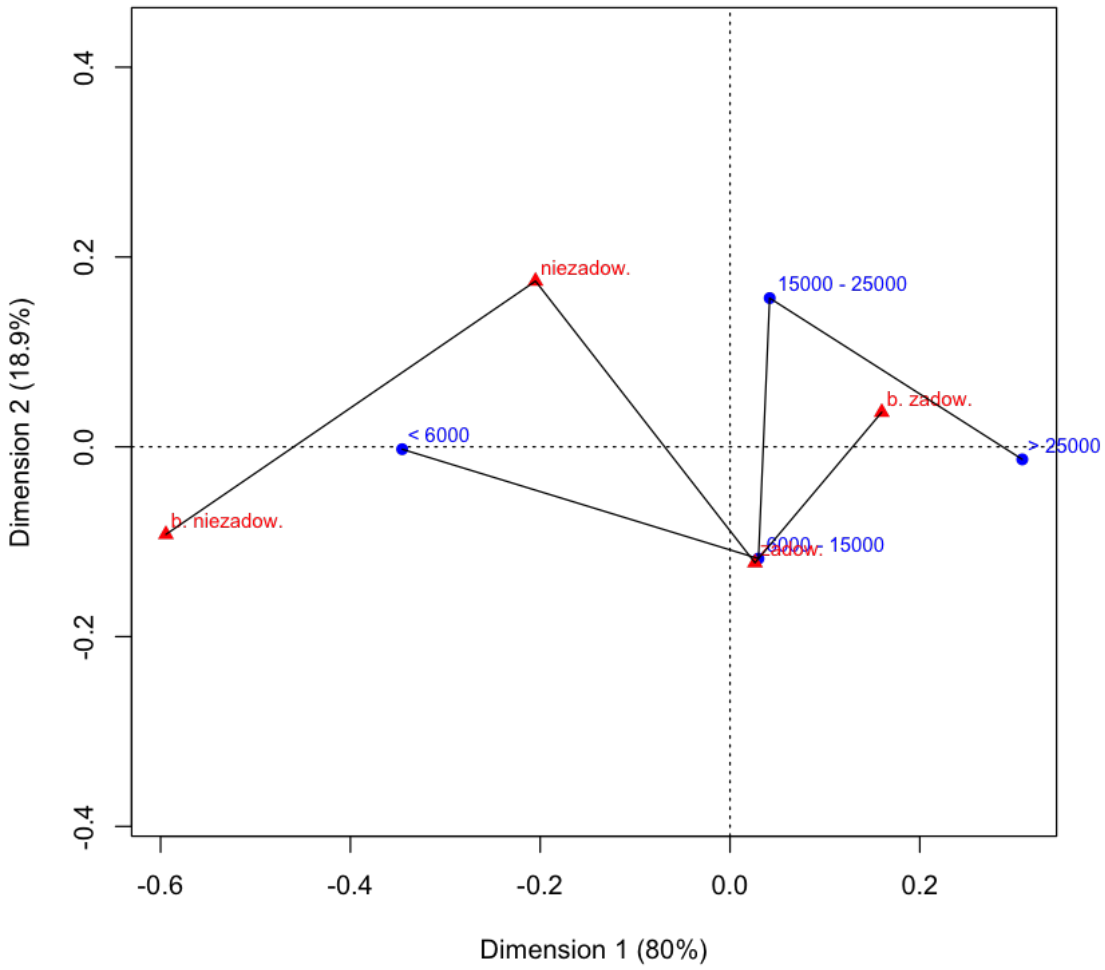
[753]:
```
sv <- svd(A)
gamm <- diag(length(sv$d)) * sv$d
U <- sv$u
V <- sv$v

F <- solve(Dr ^ (1/2)) %*% U %*% gamm
G <- solve(Dc ^ (1/2)) %*% V %*% gamm

F
```

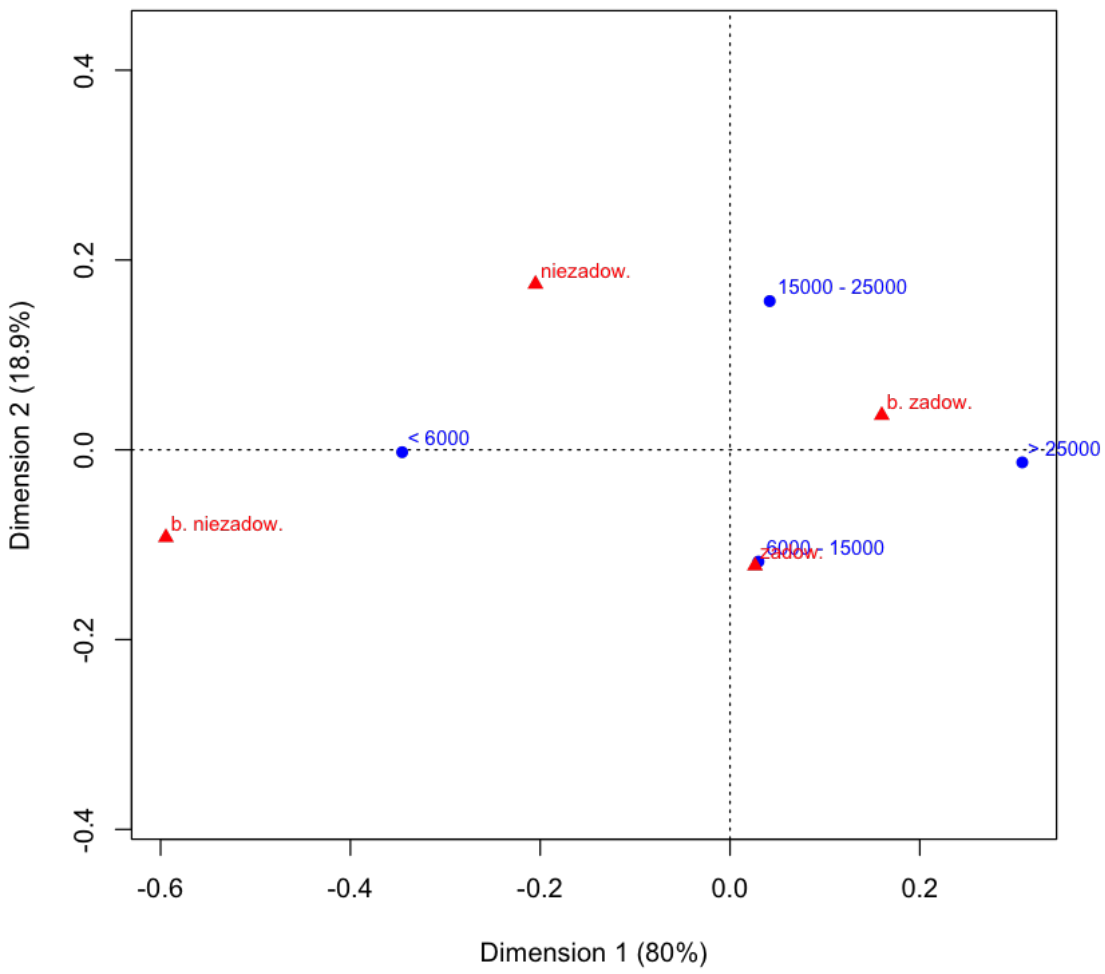
```
G
```

-0.34543922	-0.002507134	0.02175994	1.346205e-17
0.02998446	-0.117785342	-0.02235324	1.346205e-17
0.04182549	0.156690466	-0.01811548	1.346205e-17
0.30798818	-0.013250444	0.03646010	1.346205e-17

-0.5943912	-0.0923127	0.04549626	-1.346205e-17
-0.2049741	0.1746038	-0.02887773	-1.346205e-17
0.0263079	-0.1221969	-0.02260959	-1.346205e-17
0.1599100	0.0364195	0.01831000	-1.346205e-17

W ten sam sposób będziemy weryfikować poprawność naszej metody - porównując wykres wygenerowany przez wbudowaną funkcję *ca* z wykresem stworzonym przez dane pochodzące z obliczonych macierzy *F* i *G*. Punkty wygenerowane przez nas zwizulizujemy za pomocą wykresu liniowego. Wówczas punkty nie będą się pokrywać i będziemy mogli zweryfikować poprawność naszej metody.





2.3.2 Wnioski

Osoby zarabiające od 6000 do 15000 zwykle określają swój stopień zadowolenia z pracy jako "zadowolony". W tym przypadku możemy dostrzec największą zależność. Najmniejszy stopień zadowolenia deklarowały najczęściej osoby zarabiające poniżej 6000. Największa część osób, które są bardzo zadowolone z pracy, stanowiły osoby zarabiające 15000-25000 oraz powyżej 25000.