Piotr Zieleń

Sprawozdanie 2

20 lipca 2022

Spis treści

1. Wstęp

Celem sprawozdania jest przedstawienie rozwiązań oraz wniosków z rozwiązywanych podczas zajęć labolatoryjnych kolejnych list zadań.

2. Lista 5-6-7

W tej liście przeanalizuję zadania od 2. do 5., ponieważ pierwsze zadanie było jedynie ćwiczeniowe i zostało zrealizowane na zajęciach.

W pierwszym kroku zaimportowałem biblioteki, które przydały się do rozwiązywanych zadań:

```
library(binom)
library(stats)
```

2.1. Zadanie 2.

Celem zadania 2. jest przeanalizowanie wyników ankiety, którą sporządzono w kilku losowo wybranych aptekach. Klienci w różnym wieku zostali poproszeni o wskazanie najczęściej kupowanego przez nich leku przeciwbólowego. Dane zostały przestawione na tabeli (1).

	Wiek ankietowanych			
Lek	do lat 35	od 36 do 55	powyżej 55	Suma
Ibuprom	35	0	0	35
Apap	22	22	0	44
Paracetamol	15	15	15	45
Ibuprofen	0	40	10	50
Panadol	18	3	5	26
Suma	90	80	30	200

Tabela 1. Tablica dwudzielcza zależności wyboru leku, od wieku

Na podstawie tych danych, należało zweryfikować hipotezy opisane w (2.1.1), (2.1.2), (2.1.3), (2.1.4) na poziomie istotności $\alpha = 0.05$, podać wartość poziomu krytycznego i sformułować odpowiedź.

Wykonamy trzy testy: `binom.test` oraz `prop.test` z oraz bez uwzględnienia poprawki ciągłości Yatesa.

2.1.1. (a)

H_0**:** prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap jest mniejsze bądź równe $\frac{1}{4}$

H_1**:** prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap jest większe od $\frac{1}{4}$

```
apap <- 44
n <- 200
```

```
binom.test(apap, n, p=1/4, alternative="g")$p.value
```

```
## [1] 0.8562401
```

```r
prop.test(apap, n, p=1/4, alternative="g", correct=T)$p.value
```

```
## [1] 0.8154462
```

```r
prop.test(apap, n, p=1/4, alternative="g", correct=F)$p.value
```

```
## [1] 0.8364066
```

W każdym z powyższych testów uzyskałem p-wartość znacznie większą niż założony poziom istotności α, więc nie mamy podstaw do odrzucenia hipotezy zerowej H_0, czyli prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap, jest mniejsze lub równe $\frac{1}{4}$.

2.1.2. (b)

H_0**:** prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap jest równe $\frac{1}{2}$

H_1**:** prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap jest różne od $\frac{1}{2}$

```r
apap <- 44
n <- 200
```

```r
binom.test(apap, n, p=1/2, alternative="t")$p.value
```

```
## [1] 6.837911e-16
```

```r
prop.test(apap, n, p=1/2, alternative="t", correct=T)$p.value
```

```
## [1] 4.197514e-15
```

```r
prop.test(apap, n, p=1/2, alternative="t", correct=F)$p.value
```

```
## [1] 2.382836e-15
```

W każdym z przeprowadzonych testów uzyskałem p-wartość praktycznie równą zero, więc możemy założyć, że hipoteza zerowa H_0 jest nieprawdziwa, czyli zakładamy, że prawdziwa jest hipoteza alternatywna H_1, czyli prawdopodobieństwo, że losowo wybrana osoba z badanej populacji, w przypadku bólu zażywa Apap, jest różne od $\frac{1}{2}$.

2.1.3. (c)

H_0**:** prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Ibuprom jest większe bądź równe $\frac{1}{5}$

H_1**:** prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Ibuprom jest mniejsze od $\frac{1}{5}$

```r
ibuprom <- 35
n <- 200
```

```r
binom.test(ibuprom, n, p=1/5, alternative="l")$p.value
```

```
## [1] 0.215066
```

```r
prop.test(ibuprom, n, p=1/5, alternative="l", correct=T)$p.value
```

```
## [1] 0.2131628
```

```r
prop.test(ibuprom, n, p=1/5, alternative="l", correct=F)$p.value
```

```
## [1] 0.1883796
```

W każdym z przeprowadzonych testów uzyskałem p-wartość większą niż założony poziom istotności α, więc możemy założyć, że hipoteza zerowa H_0 jest prawdziwa, czyli prawdopodobieństwo, ze losowo wybrana osoba z badanej populacji, w przypadku bólu zażywa Ibuprom, jest większe bądź równe $\frac{1}{5}$.

2.1.4. (d)

W podpunkcie (d) powtarzamy pozostałe podpunkty, ale bierzemy pod uwagę tylko grupę wiekową do lat 35.

(a)

H_0**:** prawdopodobieństwo, że losowo wybrana osoba, do 35 lat, z badanej populacji w przypadku bólu zażywa Apap jest mniejsze bądź równe $\frac{1}{4}$

H_1**:** prawdopodobieństwo, że losowo wybrana osoba, do 35 lat, z badanej populacji w przypadku bólu zażywa Apap jest większe od $\frac{1}{4}$

```r
apap_35 <- 22
n_35 <- 90
```

```r
binom.test(apap_35, n_35, p=1/4, alternative="g")$p.value
```

```
## [1] 0.5885826
```

```r
prop.test(apap_35, n_35, p=1/4, alternative="g",
                                correct=T)$p.value
```

```
## [1] 0.5
```

```r
prop.test(apap_35, n_35, p=1/4, alternative="g",
                                correct=F)$p.value
```

```
## [1] 0.5484381
```

(b)

H_0**:** prawdopodobieństwo, że losowo wybrana osoba, do 35 lat, z badanej populacji w przypadku bólu zażywa Apap jest równe $\frac{1}{2}$

H_1**:** prawdopodobieństwo, że losowo wybrana osoba, do 35 lat, z badanej populacji w przypadku bólu zażywa Apap jest różne od $\frac{1}{2}$

```r
apap_35 <- 22
n_35 <- 90
```

```r
binom.test(apap_35, n_35, p=1/2, alternative="t")$p.value
```

```
## [1] 1.249258e-06
```

```r
prop.test(apap_35, n_35, p=1/2, alternative="t",
                                correct=T)$p.value
```

```
## [1] 2.101436e-06
```

```r
prop.test(apap_35, n_35, p=1/2, alternative="t",
                                correct=F)$p.value
```

```
## [1] 1.241945e-06
```

(c)

H_0**:** prawdopodobieństwo, że losowo wybrana osoba, do 35 lat, z badanej populacji w przypadku bólu zażywa Ibuprom jest większe bądź równe $\frac{1}{5}$

H_1**:** prawdopodobieństwo, że losowo wybrana osoba, do 35 lat, z badanej populacji w przypadku bólu zażywa Ibuprom jest mniejsze od $\frac{1}{5}$

```r
ibuprom_35 <- 35
n_35 <- 90
```

```r
binom.test(ibuprom_35, n, p=1/5, alternative="l")$p.value
```

```
## [1] 0.215066
```

```r
prop.test(ibuprom_35, n, p=1/5, alternative="l",
                                correct=T)$p.value
```

```
## [1] 0.2131628
```

```r
prop.test(ibuprom_35, n, p=1/5, alternative="l",
                                correct=F)$p.value
```

```
## [1] 0.1883796
```

Dla każdego z rozważanych przypadków — **(a)**, **(b)**, **(c)** uzyskałem p – wartości, dla których możemy wyciągnąć podobne wnioski, jak dla badania dla całej badanej populacji.

- W przypadku **(a)** nie ma podstaw do odrzucenia hipotezy zerowej H_0,
- W przypadku **(b)** możemy założyć, że hipoteza zerowa H_0 jest fałszywa, więc przyjmujemy hipotezę alternatywaną H_1
- W przypadku **(c)** nie ma podstaw do odrzucenia hipotezy zerowej H_0.

2.2. Zadanie 3.

Celem zadania jest weryfikacja hipotezy na poziomie istotności $\alpha = 0.05$, że prawdopodobieństwo, że osoba do lat 35 zażywa „Panadol" jest równe prawdopodobieństwu, że osoba od 36 lat do 55 lat zażywa „Panadol". W zadaniu wykorzystam test Fishera oraz opierać się będę na danych z tabeli (1).

Następnie postaram się odpowiedzieć na pytanie, czy na podstawie uzyskanego wyniku można odrzucić hipotezę o niezależności wyboru leku „Panadol" w leczeniu bólu od wieku, przy uwzględnieniu tylko dwóch grup wiekowych wymienionych wyżej.

W pierwszym kroku przedstawię tablicę dwudzielczą, dla danych — tabela (2):

```r
data <- matrix(c(18, 72, 3, 77), nrow=2, ncol=2)
```

	Wiek ankietowanych	
	do lat 35	od 36 do 55 lat
Panadol	18	3
Inny lek	72	77

Tabela 2. Zależność wyboru leku „Panadol" lub innego leku, od wieku

Hipotezy testu Fishera:

- H_0: prawdopodobieństwo, że osoba do lat 35 w przypadku bólu zażywa „Panadol" jest równe prawdopodobieństwu, że osoba od 36 do 55 lat zażywa „Panadol" w przypadku bólu
- H_1: prawdopodobieństwo, że osoba do lat 35 w przypadku bólu zażywa „Panadol" jest nie jest równe prawdopodobieństwu, że osoba od 36 do 55 lat zażywa „Panadol" w przypadku bólu

```r
fisher.test(data)$p.value
```

```
## [1] 0.001788538
```

Uzyskałem p-wartość mniejszą niż założony poziom istotności $\alpha = 0.05$, więc są podstawy, aby odrzucić hipotezę zerową i przyjąć hipotezę alternatywną, czyli zakładamy, że prawdopodobieństwo, że osoba do lat 35 w przypadku bólu zażywa „Panadol" jest nie jest równe prawdopodobieństwu, że osoba od 36 do 55 lat zażywa „Panadol" w przypadku bólu.

Można też na podstawie tego wyniku odrzucić hipotezę o niezależności (na podstawie równoważności hipotez) wyboru leku „Panadol" w leczeniu bólu od wieku, przy uwzględnieniu tylko dwóch grup wiekowych — do lat 35 i od 36 do 55 lat.

2.3. Zadanie 4.

W zadaniu 4. należało zweryfikować hipotezę o niezależności stopnia zadowolenia z pracy i wynagrodzenia, korzystając z funkcji `chisq.test`. Dane do przeanalizowania przedstawiłem w tabeli (3). Zadanie dotyczy weryfikacji hipotezy zerowej na poziomie istotności $\alpha = 0.05$.

	Stopień zadowolenia z pracy				
Wynagrodzenie	b. niezadow.	niezadow.	zadow.	b. zadow.	Suma
do 6000	20	24	80	82	206
6000-15000	22	38	104	125	289
15000-25000	13	28	81	113	235
powyżej 25000	7	18	54	92	171
Suma	62	108	319	412	901

Tabela 3. Dane do zadania 4. i 5.

Hipotezy testu chi-kwadrat:

- H_0: Stopień zadowolenia z pracy i wynagrodzenie są niezależne
- H_1: Stopień zadowolenia z pracy i wynagrodzenie nie sa niezależne

```
(data2 <- matrix(c(20, 22, 13, 7,
                   24, 38, 28, 18,
                   80, 104, 81, 54,
                   82, 125, 113, 92),
                   nrow=4, ncol=4))
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   20   24   80   82
## [2,]   22   38  104  125
## [3,]   13   28   81  113
## [4,]    7   18   54   92
```

```
chisq.test(data2)$p.value
```

```
## [1] 0.2139542
```

Uzyskałem p-wartość większą niż założony poziom istotności $\alpha = 0.05$, więc nie mamy podstaw do odrzucenia hipotezy o niezależności, czyli zakładamy, że stopień zadowolenia z pracy i wynagrodzenie są niezależne.

2.4. Zadanie 5.

Celem zadania było napisanie deklaracji funkcji, która dla danych w tablicy dwudzielczej oblicza wartość poziomu krytycznego (p-value) w asymptotycznym teście niezależności opartym na ilorazie wiarogodności, dla danych z tabeli (3). Dodatkowo należało podać hipotezy dla asymptotycznego testu niezależności, opartym na ilorazie wiarogodności, wyznaczyć p-wartość na podstawie napisanej funkcji i wyciągnąć wniosek.

Hipotezy dla asymptotycznego testu niezależności:

- H_0: Stopień zadowolenia z pracy i wynagrodzenie są niezależne
- H_1: Stopień zadowolenia z pracy i wynagrodzenie nie sa niezależne

```
(Data <- matrix(c(20, 22, 13, 7,
                  24, 38, 28, 18,
                  80, 104, 81, 54,
                  82, 125, 113, 92),
        nrow=4, ncol=4))
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   20   24   80   82
## [2,]   22   38  104  125
## [3,]   13   28   81  113
## [4,]    7   18   54   92
```

```
funkcja <- function(data){
        n <- sum(data)
        N <- matrix(rowSums(data), nrow=nrow(data),
                    ncol=nrow(data))
        R <- t(matrix(colSums(data), nrow=ncol(data),
                    ncol=ncol(data)))
        G2 <- prod(((N*R)/(n * data)) ^ data)
```

```
        G2 <- -2*log(G2)
        1-pchisq(G2, (length(N[1,])-1)*(length(R[1,])-1))
}
```

Następnie obliczam wartość krytyczną dla danych z zadania 4. — tabela (3), korzystając z napisanej funkcji:

```
funkcja(Data)
```

```
## [1] 0.2112391
```

Podobnie jak w teście chi-kwadrat – uzyskałem p-wartość większą niż zadany poziom istotności $\alpha = 0.05$, więc nie ma podstaw do odrzucenia hipotezy zerowej — można założyć, że stopień zadowolenia z pracy i wynagrodzenie są niezależne.

3. Lista 8_9

Zaimportowałem biblioteki, które przydały się do rozwiązania zadań:

```
library(stats)
library(DescTools)
library(psych)
library(expm)
library(matlib)
library(ca)
```

Dodatkowo, w każdym z poniższych zadań będę przeprowadzać analizę korespondencji, dla której kod zamieszczam poniżej:

```
analiza.kores <- function(Dane, title_='', columnsName='',
                          rowsName='', dispMatrix=FALSE) {
  nc <-  colSums(Dane)
  nr <- rowSums(Dane)
  n <- sum(nr)
  P <- Dane / n
  c <- nc / n
  r <- nr / n
  Dc <- diag(c)
  Dr <- diag(r)
  A <-inv(Dr ^ (1/2)) %*% (P- r %*% t(c)) %*% inv(Dc ^ (1/2))
  U <-  svd(A)$u
  V <-  svd(A)$v
  Gamma <-  diag(svd(A)$d)
  dispF <- inv(Dr^(1/2)) %*% U %*% Gamma
  dispG <- inv(Dc^(1/2)) %*% V %*% Gamma
  F_ <- dispF[,1:2]
  G <- dispG[,1:2]
  gam <- svd(A)$d ^ 2

  # Creating labels
  if (rowsName=='') {
```

```
    rowsName <- c()
    for (i in 1:nrow(Dane)) {
      str_ = paste('row', i, sep = '')
      rowsName <- append(rowsName, str_)
    }
  }

  if (columnsName=='') {
    columnsName <- c()
    for (i in 1:ncol(Dane)) {
      str_ = paste('col', i, sep = '')
      columnsName <- append(columnsName, str_)
    }
  }
  xlabel <- paste("Dimension 1 (", round(sum(gam[1])/sum(gam), 3) * 100, '%)',
                  sep = '')
  ylabel <- paste("Dimension 2 (", round(sum(gam[2])/sum(gam), 3) * 100, '%)',
                  sep = '')
  plot(F_, col='blue', pch = 19,
       main=title_,
       xlab=xlabel,
       ylab=ylabel,
       xlim=c(min(c(min(F_), min(G[,1])))-0.01, max(c(max(F_[,1]),
                max(G[,1])))),
       ylim=c(min(c(min(F_[,2]), min(G[,2])))-0.001, max(c(max(F_[,2]),
                max(G[,2])))))
  abline(h=0)
  abline(v=0)
  text(F_, rowsName, pos=2)

  points(G, col = "red", pch = 19)
  text(G, columnsName, pos=2)
  grid(lty=6, col="grey")

  if (dispMatrix) {
    c(sum(gam), sum(gam[1:2])/sum(gam), P, A, dispF, dispG, c, r)
  }
}
```

3.1. Zadanie 1.

W tym zadaniu należało przeprowadzić test o niezależności, wyznaczyć miary współzmienności, korzystając z własnej funkcji i porównać otrzymane wyniki w funkcjami wbudowanymi w pakiecie R, oraz przeprowadzić analizę korespondencji na podstawie danych z przykładu 1. na wykładzie 7.

3.1.1. Dane z przykładu 1. z wykładu 7.

W ramach projektu „Współczesne problemy ekologiczne świata", wśród studentów polskich uczelni przeprowadzono badania ankietowe na temat ich świadomości ekologicznej.

Między innymi zadano im pytanie dotyczące segregacji śmieci, na które mogli wybrać jedną z czterech poniższych odpowiedzi:

(A) *Segreguję śmieci, ponieważ jest to korzystne dla środowiska;*
(B) *Segreguję śmieci ponieważ taki jest wymóg ustalony;*
(C) *Segreguję śmieci, ponieważ wszyscy tak robią;*
(D) *Nie segreguję śmieci.*

Poniżej przedstawiłem wyniki ankiety:

- Tablica dwudzielcza dla kategorii *Segregacja* i *Wiek*, (dla połączonych kategorii wiekowych):

```
m1a <- matrix(c(888,369,50,457,263,95,10,99,208,29,2,
                        44,78,9,0,19,1,0,0,4),
                                nrow=5,byrow=TRUE)
dimnames(m1a) <- list(c("18-25","26-35","36-45",
                        "46-59","60+"),
                        c("A","B","C","D"))
```

	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

Tabela 4. Tablica dwudzielcza dla kategorii *Segregacja* i *Wiek*, (dla połączonych kategorii wiekowych)

- Tablica dwudzielcza dla kategorii *Segregacja* i *Miejsce zamieszkania*:

```
m2a <- matrix(c(505,202,19,136,240,77,14,88,181,
                63,8,105,512,159,21,294),nrow=4,byrow=TRUE)
dimnames(m2a) <- list(c("Wies","Miasto do 20 tys.",
        "Miasto 20-50 tys.","Miasto pow. 50 tys."),
                        c("A","B","C","D"))
```

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

Tabela 5. Tablica dwudzielcza dla kategorii *Segregacja* i *Miejsce zamieszkania*

- Tablice dla „wyjściowych” kategorii wiekowych (jak w przykładzie 1. z wykładu 7.):

```
m1b <- matrix(c(888,369,50,457,263,95,10,99,208,29,2,
                44,79,9,0,23),
                nrow=4,byrow=TRUE)
```

```
dimnames(m1b) <- list(c("18-25","26-35","36-45","46+"),
                      c("A","B","C","D"))
```

	A	B	C	D
18-25	888	369	50	457
26-35	263	95	10	99
36-45	208	29	2	44
46+	79	9	0	23

Tabela 6. Tablica dwudzielcza dla kategorii *Segregacja* i *Wiek*

```
m1c <- matrix(c(888,369,50,457,263,95,10,
                99,287,38,2,67),
                nrow=3,byrow=TRUE)
dimnames(m1c) <- list(c("18-25","26-35","36+"),
                      c("A","B","C","D"))
```

	A	B	C	D
18-25	888	369	50	457
26-35	263	95	10	99
36+	287	38	2	67

Tabela 7. Tablica dwudzielcza dla kategorii *Segregacja* i *Wiek*

3.1.2. Analiza niezależności

Przeprowadziłem po jednym teście niezależności dla danych.
Hipotezy dla testów statystycznych, w których badamy niezależność wieku, od segregacji:

- H_0: *Segregacja* i *Wiek* są niezależne
- H_1: *Segregacja* i *Wiek* nie są niezależne

Hipotezy dla testu, w którym badamy niezależność segregacji i miejsca zamieszkania:

- H_0: *Segregacja* i *Miejsce zamieszkania* są niezależne
- H_1: *Segregacja* i *Miejsce zamieszkania* nie są niezależne

```
fisher.test(m1a, simulate.p.value = TRUE)$p.value
```

```
## [1] 0.0004997501
```

```
chisq.test(m1a)$p.value
```

```
## [1] 9.14123e-13
```

```
fisher.test(m2a, simulate.p.value = TRUE)$p.value
```

```
## [1] 0.0004997501
```

```
chisq.test(m2a)$p.value
```

```
## [1] 1.196423e-10
```

```
fisher.test(m1b, simulate.p.value = TRUE)$p.value
```

```
## [1] 0.0004997501
```

```
chisq.test(m1b)$p.value
```

```
## [1] 3.940232e-12
```

```
fisher.test(m1c, simulate.p.value = TRUE)$p.value
```

```
## [1] 0.0004997501
```

```
chisq.test(m1c)$p.value
```

```
## [1] 1.874402e-13
```

W każdym z poniższych testów, otrzymałem p-wartość mniejszą niż założony poziom istotności $\alpha = 0.05$, więc można założyć, że hipoteza zerowa H_0 jest fałszywa i ją odrzucić, a przyjąć hipotezę alternatywną H_1, czyli w każdym przypadku zakładamy, że miejsce zamieszkania i segregacja śmieci nie są niezależne. Widać jednak, że dla każdego z przeprowadzonych testów, p-wartość testu Fishera wynosi dokładnie tyle samo. Można na tej podstawie sądzić, że nie działa on dobrze, dla naszych danych, być może przez niespełnione niektóre założenia.

3.1.3. Funkcja do wyliczania odpowiednich miar współzmienności zmiennych

Napisałem funkcję, która jako argument przyjmuje dane oraz typ miary współzmienności, którą należy wyliczyć:

```
funkcja <- function(data, method){
        R <- nrow(data)
        C <- ncol(data)
        if (method == "GoodmanKruskalTau"){
                suma1 <- sum(sum(data^2/(sum(data) *
                        rowSums(data))))
                suma2 <- sum((colSums(data)/sum(data))^2)
                (suma1 - suma2)/(1 - suma2)
        } else if (method == "CramerV") {
                sqrt((as.numeric(chisq.test(data)$statistic))
                        /(sum(data) * min(R - 1, C - 1)))
        } else if (method == "T-Czuprowa"){
                sqrt((as.numeric(chisq.test(data)$statistic))
                        /(sum(data) * sqrt((R-1)*(C-1))))
        } else if (method == "wspolczynnik_fi"){
                sqrt((as.numeric(chisq.test(data)$statistic))
                        /sum(data))
        } else if (method == "CPearsona"){
                sqrt((as.numeric(chisq.test(data)$statistic))
                /((as.numeric(chisq.test(data)$statistic))
                +sum(data)))
        }
}
```

3.1.4. Porównanie wyników z napisanej funkcji, z funkcjami wbudowanymi w R

Porównałem wyniki otrzymane z naszej funkcji, z funkcjami wbudowanymi w R.

— Dla danych `m1a`:

- Współczynnik Goodmana i Kruskala:

```
funkcja(m1a, "GoodmanKruskalTau")
## [1] 0.01712163
GoodmanKruskalTau(m1a, direction="column")
## [1] 0.01712163
```

- Współczynnik V Cramera:

```
funkcja(m1a, "CramerV")
## [1] 0.1029241
CramerV(m1a)
## [1] 0.1029241
```

- Współczynnik T-Czurpowa

```
funkcja(m1a, "T-Czuprowa")
## [1] 0.09578165
TschuprowT(m1a)
## [1] 0.09578165
```

- Współczynnik φ:

```
funkcja(m1a, "wspolczynnik_fi")
## [1] 0.1782697
Phi(m1a)
## [1] 0.1782697
```

- Współczynnik C Pearsona:

```
funkcja(m1a, "CPearsona")
## [1] 0.1755028
ContCoef(m1a)
## [1] 0.1755028
```

— Dla danych `m2a`:

- Współczynnik Goodmana i Kruskala:

```
funkcja(m2a, "GoodmanKruskalTau")
## [1] 0.01012792
GoodmanKruskalTau(m2a, direction="column")
## [1] 0.01012792
```

- Współczynnik V Cramera:

```
funkcja(m2a, "CramerV")
## [1] 0.09116024
CramerV(m2a)
## [1] 0.09116024
```

- Współczynnik T-Czurpowa

```
funkcja(m2a, "T-Czuprowa")
## [1] 0.09116024
TschuprowT(m2a)
## [1] 0.09116024
```

- Współczynnik φ:

```
funkcja(m2a, "wspolczynnik_fi")
## [1] 0.1578942
Phi(m2a)
## [1] 0.1578942
```

- Współczynnik C Pearsona:

```
funkcja(m2a, "CPearsona")
## [1] 0.155962
ContCoef(m2a)
## [1] 0.155962
```

— Dla danych `m1b`:

- Współczynnik Goodmana i Kruskala:

```
funkcja(m1b, "GoodmanKruskalTau")
## [1] 0.01508301
GoodmanKruskalTau(m1b, direction="column")
## [1] 0.01508301
```

- Współczynnik V Cramera:

```
funkcja(m1b, "CramerV")
## [1] 0.09627191
CramerV(m1b)
## [1] 0.09627191
```

- Współczynnik T-Czurpowa

```
funkcja(m1b, "T-Czuprowa")
## [1] 0.09627191
TschuprowT(m1b)
## [1] 0.09627191
```

- Współczynnik φ:

```
funkcja(m1b, "wspolczynnik_fi")
## [1] 0.1667478
Phi(m1b)
## [1] 0.1667478
```

- Współczynnik C Pearsona:

```
funkcja(m1b, "CPearsona")
## [1] 0.1644769
ContCoef(m1b)
## [1] 0.1644769
```

— Dla danych `m1c`:

- Współczynnik Goodmana i Kruskala:

```
funkcja(m1c, "GoodmanKruskalTau")
## [1] 0.01489645
GoodmanKruskalTau(m1c, direction="column")
## [1] 0.01489645
```

- Współczynnik V Cramera:

```
funkcja(m1c, "CramerV")
## [1] 0.1168235
CramerV(m1c)
## [1] 0.1168235
```

- Współczynnik T-Czurpowa

```
funkcja(m1c, "T-Czuprowa")
## [1] 0.1055619
TschuprowT(m1c)
## [1] 0.1055619
```

- Współczynnik φ:

```
funkcja(m1c, "wspolczynnik_fi")
## [1] 0.1652133
Phi(m1c)
## [1] 0.1652133
```

- Współczynnik C Pearsona:

```
funkcja(m1c, "CPearsona")
## [1] 0.1630037
ContCoef(m1c)
## [1] 0.1630037
```

3.1.5. Analiza Korespondencji

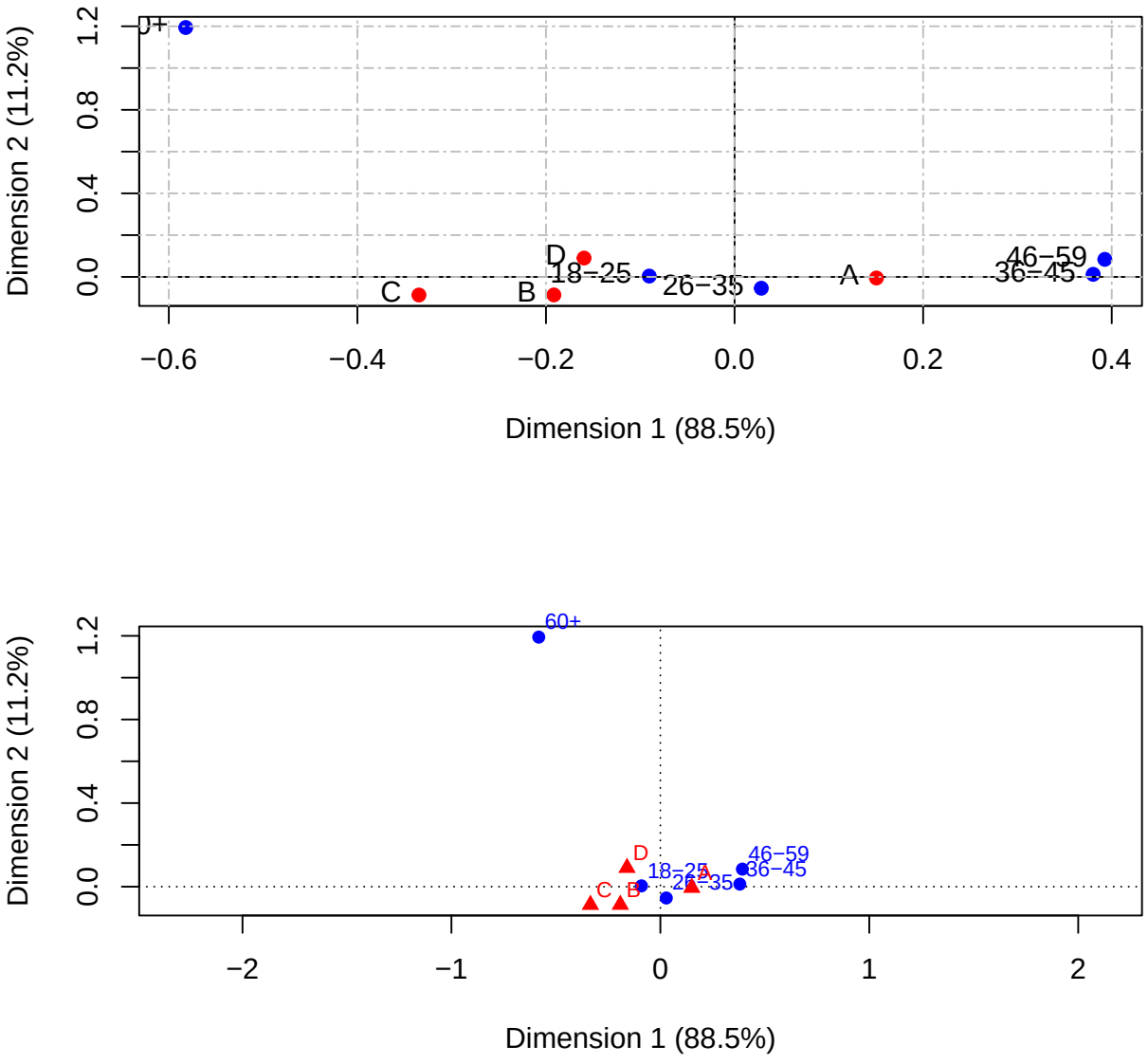
Poniżej przedstawiłem macierze oraz wykresy, które uzyskałem z naszej funkcji, w celu przeprowadzenia analizy korespondencji. Wartości w macierzy zaokrągliłem do czwartego miejsca po przecinku:

— Dla danych `m1a`:

$$\mathbf{P:}\begin{bmatrix} 0.3383 & 0.1406 & 0.019 & 0.1741 \\ 0.1002 & 0.0362 & 0.0038 & 0.0377 \\ 0.0792 & 0.011 & 8\times10^{-4} & 0.0168 \\ 0.0297 & 0.0034 & 0 & 0.0072 \\ 4\times10^{-4} & 0 & 0 & 0.0015 \end{bmatrix},\mathbf{A:}\begin{bmatrix} -0.0492 & 0.0336 & 0.0252 & 0.0366 \\ 0.0088 & 0.0118 & -0.0061 & -0.0219 \\ 0.083 & -0.0666 & -0.0354 & -0.0552 \\ 0.0511 & -0.0489 & -0.0309 & -0.024 \\ -0.0205 & -0.0191 & -0.0067 & 0.0504 \end{bmatrix} \tag{1}$$

$$\mathbf{R:}\begin{bmatrix} 0.672 & 0.1779 & 0.1078 & 0.0404 & 0.0019 \end{bmatrix},\mathbf{C:}\begin{bmatrix} 0.5478 & 0.1912 & 0.0236 & 0.2373 \end{bmatrix} \tag{2}$$

$$\mathbf{F:}\begin{bmatrix} -0.0905 & 0.0039 & -0.0042 & 0 \\ 0.0284 & -0.0542 & 0.0195 & 0 \\ 0.3804 & 0.0123 & -0.0129 & 0 \\ 0.3927 & 0.084 & 0.0158 & 0 \\ -0.582 & 1.1941 & 0.0714 & 0 \end{bmatrix},\mathbf{G:}\begin{bmatrix} 0.1505 & -0.0052 & -0.0012 & 0 \\ -0.1916 & -0.0863 & 0.0098 & 0 \\ -0.3347 & -0.0867 & -0.0643 & 0 \\ -0.1596 & 0.0902 & 0.0013 & 0 \end{bmatrix} \tag{3}$$



Rysunek 1. Porównanie wykresu analizy korespondencji z naszej funkcji (na górze) z wykresem funkcji wbudowanej (na dole), dla danych m1a

— Dla danych `m2a`:

$$\mathbf{P:}\begin{bmatrix}0.1925 & 0.077 & 0.0072 & 0.0518\\ 0.0915 & 0.0293 & 0.0053 & 0.0335\\ 0.069 & 0.024 & 0.003 & 0.04\\ 0.1951 & 0.0606 & 0.008 & 0.112\end{bmatrix},\mathbf{A:}\begin{bmatrix}0.0293 & 0.0569 & -0.0059 & -0.0937\\ 0.0134 & -0.0065 & 0.0254 & -0.0225\\ -0.0204 & -0.0122 & -0.0029 & 0.0429\\ -0.0238 & -0.0416 & -0.0093 & 0.0764\end{bmatrix} \tag{4}$$

$$\mathbf{R:}\begin{bmatrix}0.3285 & 0.1597 & 0.1361 & 0.3758\end{bmatrix},\mathbf{C:}\begin{bmatrix}0.548 & 0.1909 & 0.0236 & 0.2374\end{bmatrix} \tag{5}$$

$$\mathbf{F:}\begin{bmatrix}-0.1971 & 0.0211 & 0.0015 & 0\\ -0.052 & -0.0768 & -0.0034 & 0\\ 0.1293 & 0.0224 & -0.0232 & 0\\ 0.1476 & 0.0061 & 0.0086 & 0\end{bmatrix},\mathbf{G:}\begin{bmatrix}-0.059 & -0.0124 & 0.0075 & 0\\ -0.1584 & 0.0424 & -0.013 & 0\\ -0.0355 & -0.1732 & -0.0401 & 0\\ 0.267 & 0.0116 & -0.0029 & 0\end{bmatrix} \tag{6}$$

— Dla danych `m1b`:

$$\mathbf{P:}\begin{bmatrix}0.3383 & 0.1406 & 0.019 & 0.1741\\ 0.1002 & 0.0362 & 0.0038 & 0.0377\\ 0.0792 & 0.011 & 8\times10^{-4} & 0.0168\\ 0.0301 & 0.0034 & 0 & 0.0088\end{bmatrix},\mathbf{A:}\begin{bmatrix}-0.0492 & 0.0336 & 0.0252 & 0.0366\\ 0.0088 & 0.0118 & -0.0061 & -0.0219\\ 0.083 & -0.0666 & -0.0354 & -0.0552\\ 0.0455 & -0.0518 & -0.0316 & -0.0127\end{bmatrix} \tag{7}$$

$$\mathbf{R:}\begin{bmatrix}0.672 & 0.1779 & 0.1078 & 0.0423\end{bmatrix},\mathbf{C:}\begin{bmatrix}0.5478 & 0.1912 & 0.0236 & 0.2373\end{bmatrix} \tag{8}$$

$$\mathbf{F:}\begin{bmatrix}-0.0899 & 0.0113 & -0.0027 & 0\\ 0.0241 & -0.058 & 0.0129 & 0\\ 0.3801 & -0.0156 & -0.016 & 0\\ 0.3582 & 0.1039 & 0.029 & 0\end{bmatrix},\mathbf{G:}\begin{bmatrix}0.1446 & -0.0049 & -0.0013 & 0\\ -0.2034 & -0.0495 & 0.0075 & 0\\ -0.3459 & -0.0212 & -0.0597 & 0\\ -0.1353 & 0.0533 & 0.0029 & 0\end{bmatrix} \tag{9}$$

— Dla danych `m1c`:

$$\mathbf{P:}\begin{bmatrix}0.3383 & 0.1406 & 0.019 & 0.1741\\ 0.1002 & 0.0362 & 0.0038 & 0.0377\\ 0.1093 & 0.0145 & 8\times10^{-4} & 0.0255\end{bmatrix},\mathbf{A:}\begin{bmatrix}-0.0492 & 0.0336 & 0.0252 & 0.0366\\ 0.0088 & 0.0118 & -0.0061 & -0.0219\\ 0.0945 & -0.084 & -0.0467 & -0.0535\end{bmatrix} \tag{10}$$

$$\mathbf{R:}\begin{bmatrix}0.672 & 0.1779 & 0.1501\end{bmatrix},\mathbf{C:}\begin{bmatrix}0.5478 & 0.1912 & 0.0236 & 0.2373\end{bmatrix} \tag{11}$$

$$\mathbf{F:}\begin{bmatrix}0.0899 & 0.0119 & 0\\ -0.024 & -0.0595 & 0\\ -0.374 & 0.0173 & 0\end{bmatrix},\mathbf{G:}\begin{bmatrix}-0.1444 & -0.0032 & 0\\ 0.204 & -0.0437 & 0\\ 0.3472 & 0.0197 & 0\\ 0.1345 & 0.0406 & 0\end{bmatrix} \tag{12}$$

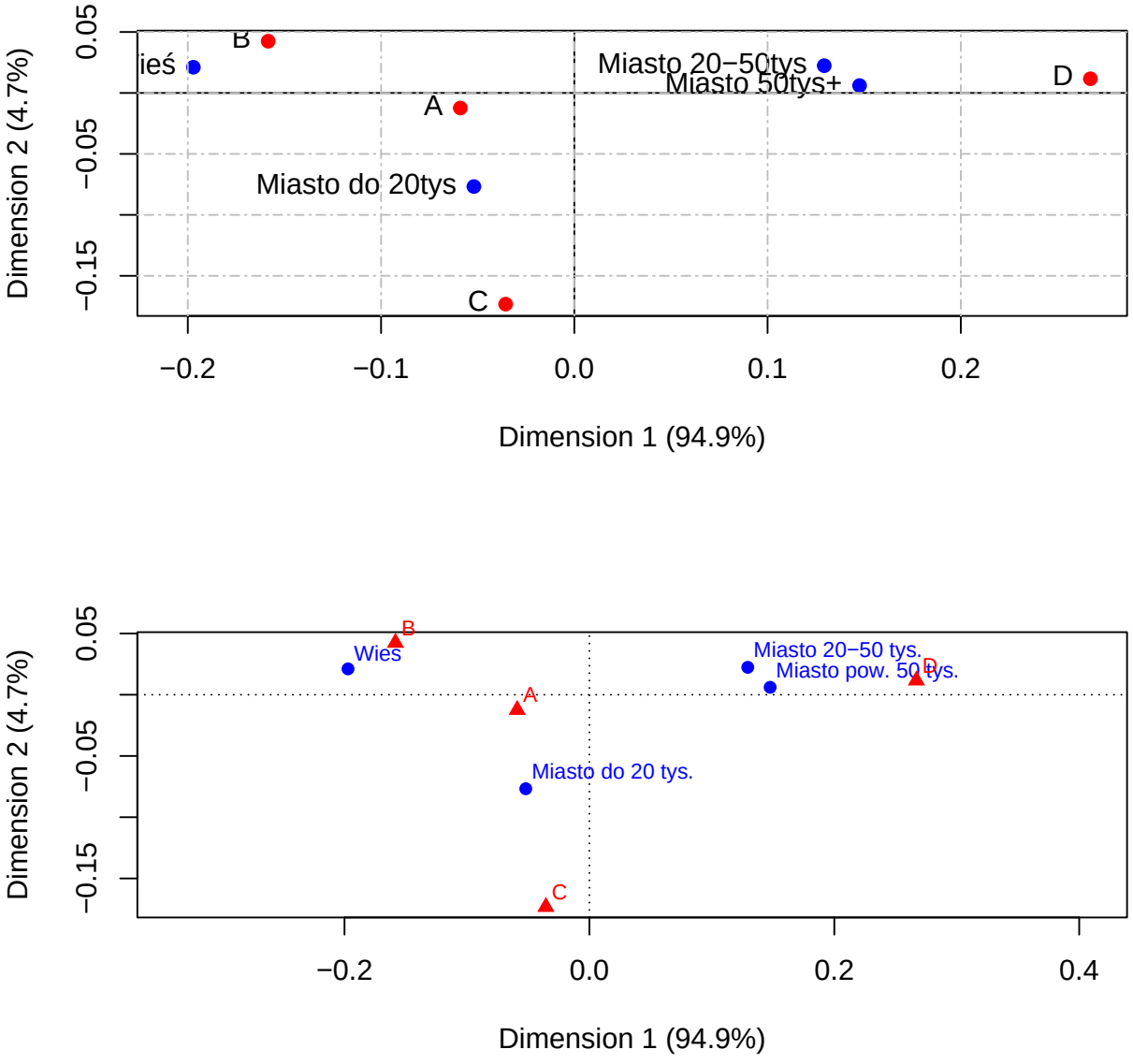
Wyznaczyłem jeszcze inercję całkowitą dla wszytskich danych, korzystając ze wzoru:

$$\lambda = \mathrm{tr}\left(\mathbf{A}^T\mathbf{A}\right) \tag{13}$$

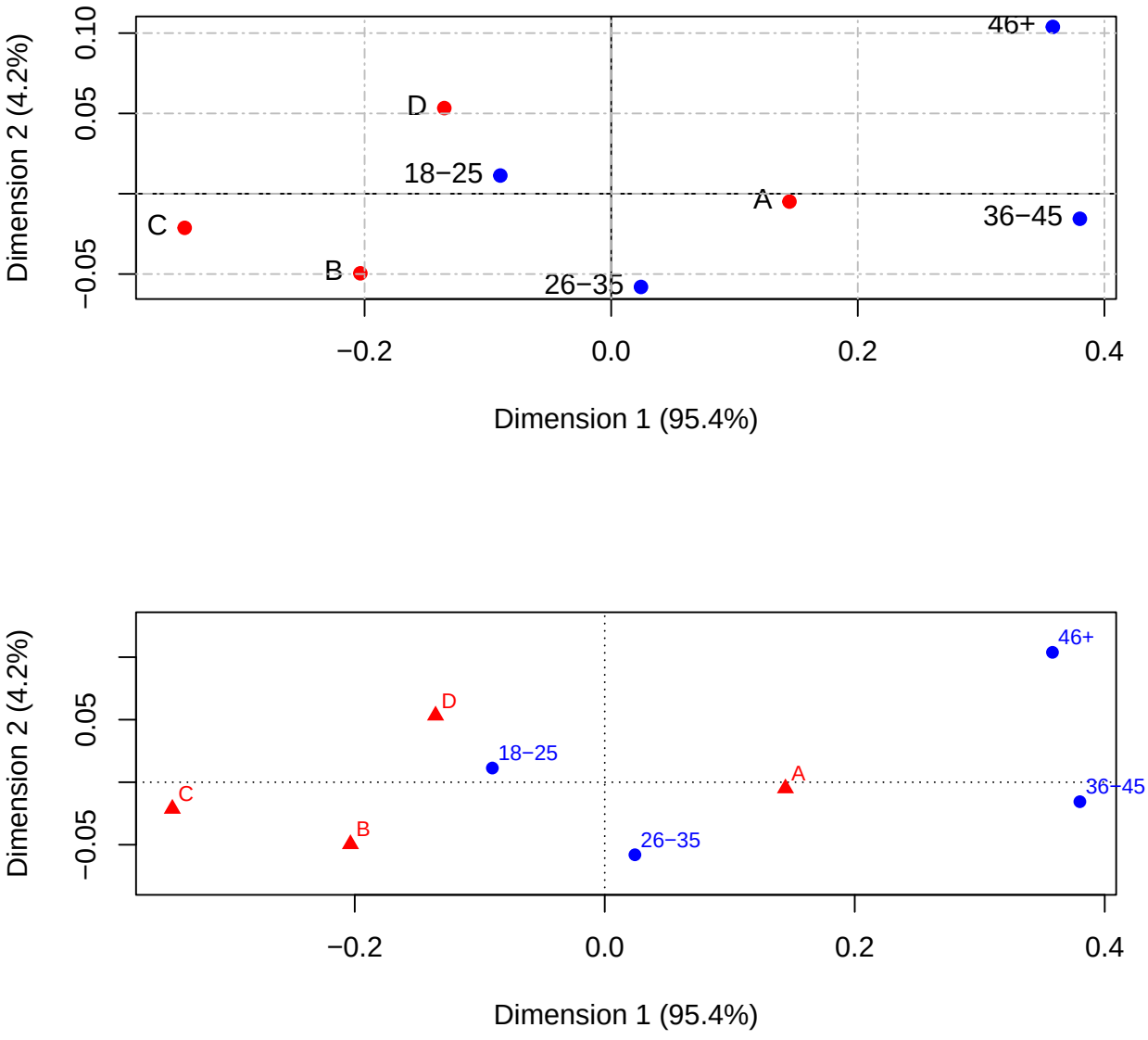
gdzie macierz $\mathbf{A}$, to macierz uzyskana podczas analizy korespondencji i otrzymałem następujące wartości:

- Dla danych `m1a`:

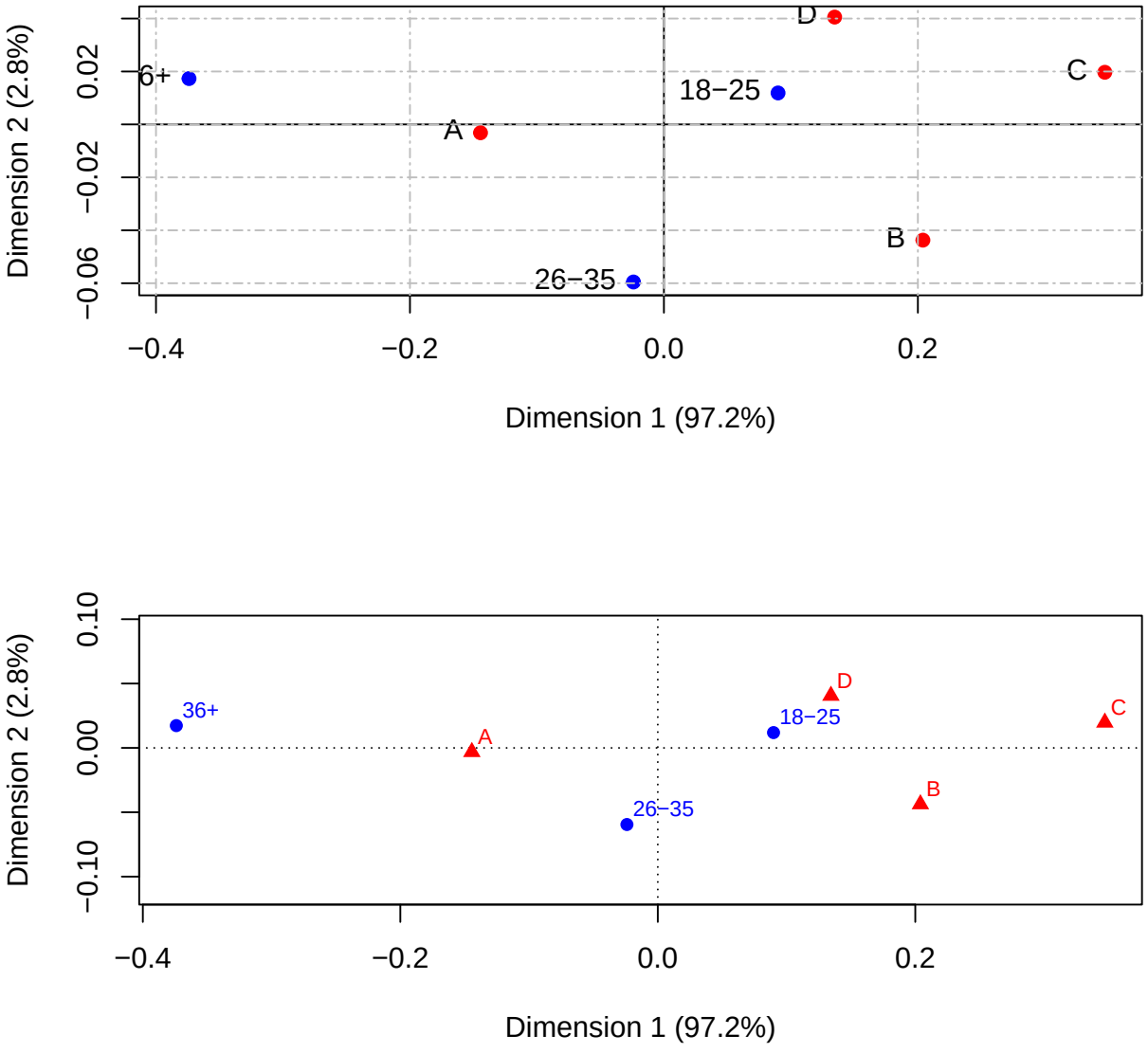
```
m1aA

##               [,1]        [,2]         [,3]        [,4]
## [1,] -0.049184998  0.03363991  0.025206501  0.03657666
## [2,]  0.008753063  0.01175547 -0.006053731 -0.02194086
## [3,]  0.083034110 -0.06664708 -0.035362706 -0.05516979
## [4,]  0.051053163 -0.04886150 -0.030882999 -0.02396049
## [5,] -0.020509120 -0.01908568 -0.006707359  0.05040715
```



Rysunek 2. Porównanie wykresu analizy korespondencji z naszej funkcji (na górze) z wykresem funkcji wbudowanej (na dole), dla danych m2a



Rysunek 3. Porównanie wykresu analizy korespondencji z naszej funkcji (na górze) z wykresem funkcji wbudowanej (na dole), dla danych m1b



Rysunek 4. Porównanie wykresu analizy korespondencji z naszej funkcji (na górze) z wykresem funkcji wbudowanej (na dole), dla danych m1c

```
sum(diag(m1aA %*% t(m1aA)))
```

```
## [1] 0.0317801
```

- Dla danych `m2a`:

```
m2aA
```

```
##              [,1]         [,2]         [,3]        [,4]
## [1,]  0.02928830  0.056939447 -0.005914787 -0.09369183
## [2,]  0.01337288 -0.006546968  0.025436939 -0.02247047
## [3,] -0.02043631 -0.012205686 -0.002925313  0.04291668
## [4,] -0.02380534 -0.041626547 -0.009291271  0.07642676
```

```
sum(diag(m2aA %*% t(m2aA)))
```

```
## [1] 0.02493057
```

- Dla danych `m1b`:

```
m1bA
```

```
##               [,1]        [,2]         [,3]        [,4]
## [1,] -0.049184998  0.03363991  0.025206501  0.03657666
## [2,]  0.008753063  0.01175547 -0.006053731 -0.02194086
## [3,]  0.083034110 -0.06664708 -0.035362706 -0.05516979
## [4,]  0.045537250 -0.05179905 -0.031602979 -0.01271630
```

```
sum(diag(m1bA %*% t(m1bA)))
```

```
## [1] 0.02780484
```

- Dla danych `m1c`:

```
m1cA
```

```
##               [,1]        [,2]         [,3]        [,4]
## [1,] -0.049184998  0.03363991  0.025206501  0.03657666
## [2,]  0.008753063  0.01175547 -0.006053731 -0.02194086
## [3,]  0.094542479 -0.08397793 -0.046744450 -0.05350652
```

```
sum(diag(m1cA %*% t(m1cA)))
```

```
## [1] 0.02729545
```

3.1.6. Wnioski

Analizę rozpocząłem od przeprowadzenia testów niezależności Fishera i chi-kwadrat, na poziomie istotności $\alpha = 0.01$. Test Fishera dla każdej z tabel dwudzielczych ((4), (5), (6), (7)) wyznaczył taką samą p-wartość równą w przybliżeniu 0.0005. Przez to, że wszystkie p-wartości są dokładnie takie same, można sądzić, że nie są spełnione założenia do przeprowadzenia testu Fishera, jak na przykład to, że w niektórych pozycjach tabeli występują zera, lub bardzo małe wartości. P-wartości uzyskane w teście chi-kwadrat są praktycznie równe zero, więc są podstawy do odrzucenia hipotezy o niezależności H_0, czyli można założyć dla wszystkich zmiennych, że nie są niezależne.

Skoro odrzuciłem hipotezy o niezależności dla wszystkich rozważanych tabel dwudzielczych, to możemy teraz badać współzmienność dla zmiennych nominalnych. Współczynniki zmienności

wyznaczyłem na podstawie naszej funkcji i funkcji wbudowanej w pakiecie R. Dla wszystkich przypadków uzyskałem dokładnie takie same wartości dla funkcji wbudowanych i naszej funkcji, więc można stwierdzić, że funkcja została zaimplementowana poprawnie.

Następnie przeszedłem do analizy korespondencji, dzięki której możemy odczytać jakie zachodzą relacje między rozważanymi zmiennymi. Wykresy napisanej funkcji i wykresy funkcji wbudowanej w pakiecie R są takie same, więc można stwierdzić, że funkcja do analizy korespondencji została zaimplementowana poprawnie. Na sam koniec wyliczyłem inercję całkowitą. Inercję całkowitą możemy interpretować jako miarę rozproszenia profili w przestrzeni wielowymiarowej (im jest większa, tym punkty bardziej są rozproszone).

Poniżej przedstawiłem wnioski jakie możemy wyciągnąć dla każdych danych, na podstawie wyznaczonych współczynników zmienności i analizy korespondencji:

- Dla danych `m1a`:
  Wszystkie wyznaczone miary współzmienności mieszczą się w przedziale $[0.09, 0.18]$, przez co możemy stwierdzić, że nasze zmienne charakretyzują się słabą współzmiennością.
  Na rysunku (1) przedstawiłem dwa wykresy analizy korespondencji. Pierwsza zmienna, która się wyróżnia na wykresie, to zmienna opisująca osoby 60+, które brały udział w ankiecie. W ankiecie brało tylko 5 osób 60+, więc trudno wyciągać wnioski dla tej klasyfikacji, dlatego punkt reprezentujący te osoby znajduje się najdalej od początku układu współrzędnych. Oprócz tego możemy stwierdzić, że stosunkowo najczęstszą odpowiedzią jakie udzielały osoby w wieku $46-59$, $36-45$ i $26-35$ to odpowiedź A (*Segreguję śmieci, ponieważ jest to korzystne dla środowiska*), ponieważ punkty odpowiadające ich kategoriom, znajdują się najbliżej tej odpowiedzi. Odpowiedź D (*Nie segreguję śmieci*) była najczęściej udzielana przez osoby w wieku $18-25$ lat. Odpowiedź C (*Segreguję śmieci, ponieważ wszyscy tak robią*) była najrzadziej wybieraną spośród wszystkich, ponieważ punkt reprezentujący tę odpowiedź znajduje się najdalej od wszystkich kategorii reprezentujących wiek osób ankietowanych;
  Procenty wyjaśnienia dla dwóch osi współrzędnych sumują się prawie do 100%, więc relacja jest dość dobrze przedstawiona.
  Obliczając inercję całkowitą uzyskałem wartość w przybliżeniu równą 0.032.
- Dla danych `m2a`:
  Wszystkie wyznaczone miary współzmienności, poza współczynnikiem Goodmana i Kruskala mieszczą się przedziale $[0.09, 0.16]$, natomiast współczynnik Goodmana i Kruskala wyszedł w przybliżeniu równy 0.01. Korzystając z wykładu, wiemy, że współczynnik Goodmana i Kruskala jest najlepszy spośród rozważanych. Na tej podstawie możemy wyciągnąć wniosek, że nasze zmienne charakteryzują się bardzo słabą współzmiennością.
  Na rysunku (2) przedstawiłem wykresy analizy korespondencji. Procenty objaśnienia sumjują się prawie do 100%, więc relacja między zmiennymi jest dobrze przedstawiona. Na podstawie wykresu możemy powiedzieć, że stosunkowo, wśród ankietowanych osób, te które nie segregują śmieci mieszkają najczęściej w większych miastach (miasta powyżej 20 tys mieszkańców), ponieważ punkt reprezentujący odpowiedź D, znajduje się najbliżej punktów reprezentujących te miasta. Ludzie segregujący śmieci ze względu na taki wymóg ustawowy, stosunkowo najczęściej mieszkają na wsi. Najrzadziej wybieraną odpowiedzią, była odpowiedź C, ponieważ znajduje się na wykresie stosunkowo najdalej od innych odpowiedzi.
  Obliczając inercję całkowitą uzyskałem wartość w przybliżeniu równą 0.025. Jest mniejsza niż w przypadku danych `m1a`, co zgadza się z intuicją, bo porównując rysunki ((1), (2)) widać, że dla danych `m2a` nie ma wartości tak odstającej jak dla zmiennej `m1a`.
- Dla danych `m1b` oraz `m1c`:
  Podsumowanie dla zmiennych `m1b` i `m1c` przeprowadziłem wspólnie, ponieważ na podstawie uzyskanych wyników możemy wyciągnąć podobne wnioski.
  Zmienne `m1b` i `m1c` różnią tym, że w zmiennej `m1c` połączone zostały kategorie wiekowe $36-$

45 i 46+ z danych `m1b` (tabele: (6) (7)).
Uzyskane współczynniki zmienności mieszczą się w przedziale [0.09, 0.17] przez co możemy wnioskować, że zmienne charakteryzują się słabą współzmiennością.
Na rysunkach (3) i (4) przedstawiłem wykresy analizy korespondencji. Pierwszą rzeczą na którą warto zwrócić uwagę jest to, że połączenie kategorii 60+ z zmiennej `m1a` do kategorii 46 – 59 spowodowało, że nie widać teraz punktu, który jest znacznie oddalony od innych. Pozostałe wnioski możemy wyciągnąć podobnie jak dla zmiennej `m1a`. Odpowiedź D była udzielana stosunkowo najczęściej przez najmłodszych ankietowanych, osoby ankietowane powyżej 25 roku życia stosunkowo najczęściej segregują śmieci, ze względu na ochronę środowiska. Odpowiedź C (*Segreguję śmieci, bo inni tak robią*) była najrzadziej wybierana.
Suma procentów zmienności daje wartość równą 100%, więc relacje między kategoriami są dobrze przedstawione.
Inercja całkowita dla rozważanych danych wyszła bardzo podobna, odpowiednio dla danych `m1b` i `m1c` — 0.028 i 0.027.

3.2. Zadanie 2

W tym zadaniu należało wyznaczyć odpowiednie miary współzmienności i przeprowadzić analizę korespondencji (podobnie jak w zadaniu 1. z tej listy — (3.1))

3.2.1. Dane do analizy

W tym zadaniu analizuję dane, które już się pojawiły w zadaniu 2. w liście 5_6_7 — (2.1), (tabela (1))

```
data <- matrix(c(35, 0, 0,
                 22, 22, 0,
                 15, 15, 15,
                 0, 40, 10,
                 18, 3, 5), byrow=T, nrow=5)
dimnames(data) <- list(c('Ibuprom', 'Apap', 'Paracetamol', 'Ibuprofen', 'Panadol'),
                 c('do lat 35', 'od 36 do 55', 'powyżej 55'))
data
```

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

3.2.2. Analiza niezależności

Przeprowadziłem dwa testy niezależności (`chisq.test` i `fisher.test`) dla danych.
Hipotezy dla testów:

- H_0: *Wiek* i *Lek* są niezależne
- H_1: *Wiek* i *Lek* nie są niezależne

```
fisher.test(data, simulate.p.value=T)$p.value
```

```
## [1] 0.0004997501
```

```
chisq.test(data)$p.value
```

```
## [1] 3.617192e-21
```

Wartość p obydwu testów jest znacznie niższa od założonego poziomu istotności $\alpha = 0.05$ więc można założyć, że hipoteza zerowa H_0 jest fałszywa — *Wiek* i *Lek* nie są niezależne.

3.2.3. Miary współzmienności

W tej części pracy wyliczyłem miary współzmienności dla rozważanych danych, korzystając z funkcji napisanej w podrozdziale (3.1.3) i porównałem wyniki z funkcjami wbudowanymi.

- Współczynnik Goodmana i Kruskala:

```
funkcja(data, "GoodmanKruskalTau")
```

```
## [1] 0.3477173
```

```
GoodmanKruskalTau(data, direction="column")
```

```
## [1] 0.3477173
```

- Współczynnik V Cramera:

```
funkcja(data, "CramerV")
```

```
## [1] 0.5361155
```

```
CramerV(data)
```

```
## [1] 0.5361155
```

- Współczynnik T-Czurpowa

```
funkcja(data, "T-Czuprowa")
```

```
## [1] 0.4508176
```

```
TschuprowT(data)
```

```
## [1] 0.4508176
```

- Współczynnik φ:

```
funkcja(data, "wspolczynnik_fi")
```

```
## [1] 0.7581819
```

```
Phi(data)
```

```
## [1] 0.7581819
```

- Współczynnik C Pearsona:

```
funkcja(data, "CPearsona")
```

```
## [1] 0.6041645
```

```
ContCoef(data)
```

```
## [1] 0.6041645
```

3.2.4. Analiza korespondencji

Poniżej przedstawiłem macierze i wykresy, które uzyskałem z naszej funkcji, w celu przeprowadzenia analizy korespondencji. Wartości w macierzy zaokrągliłem do czwartego miejsca po przecinku:

$$\mathbf{P:}\begin{bmatrix}0.175 & 0.000 & 0.000\\ 0.110 & 0.110 & 0.000\\ 0.075 & 0.075 & 0.075\\ 0.000 & 0.200 & 0.050\\ 0.090 & 0.015 & 0.025\end{bmatrix},\mathbf{A:}\begin{bmatrix}0.343 & -0.2646 & -0.162\\ 0.035 & 0.0742 & -0.1817\\ -0.0825 & -0.05 & 0.2245\\ -0.3354 & 0.3162 & 0.0645\\ 0.1302 & -0.1623 & 0.0394\end{bmatrix} \tag{14}$$

$$\mathbf{R:}\begin{bmatrix}0.175 & 0.220 & 0.225 & 0.250 & 0.130\end{bmatrix},\mathbf{C:}\begin{bmatrix}0.45 & 0.40 & 0.15\end{bmatrix} \tag{15}$$

$$\mathbf{F:}\begin{bmatrix}-1.0996 & 0.1148 & 0\\ -0.0571 & 0.4211 & 0\\ 0.1856 & -0.4806 & 0\\ 0.9225 & 0.1251 & 0\\ -0.5185 & -0.2758 & 0\end{bmatrix},\mathbf{G:}\begin{bmatrix}-0.752 & 0.0376 & 0\\ 0.6739 & 0.238 & 0\\ 0.459 & -0.7474 & 0\end{bmatrix} \tag{16}$$

Wyznaczyłem inercję całkowitą, korzystając ze wzoru (13). Uzyskałem następującą wartość:

```
dataA
```

```
##              [,1]        [,2]        [,3]
## [1,]  0.34298526 -0.26457513 -0.16201852
## [2,]  0.03496029  0.07416198 -0.18165902
## [3,] -0.08249579 -0.05000000  0.22453656
## [4,] -0.33541020  0.31622777  0.06454972
## [5,]  0.13023647 -0.16225573  0.03938632
```

```
sum(diag(dataA %*% t(dataA)))
```

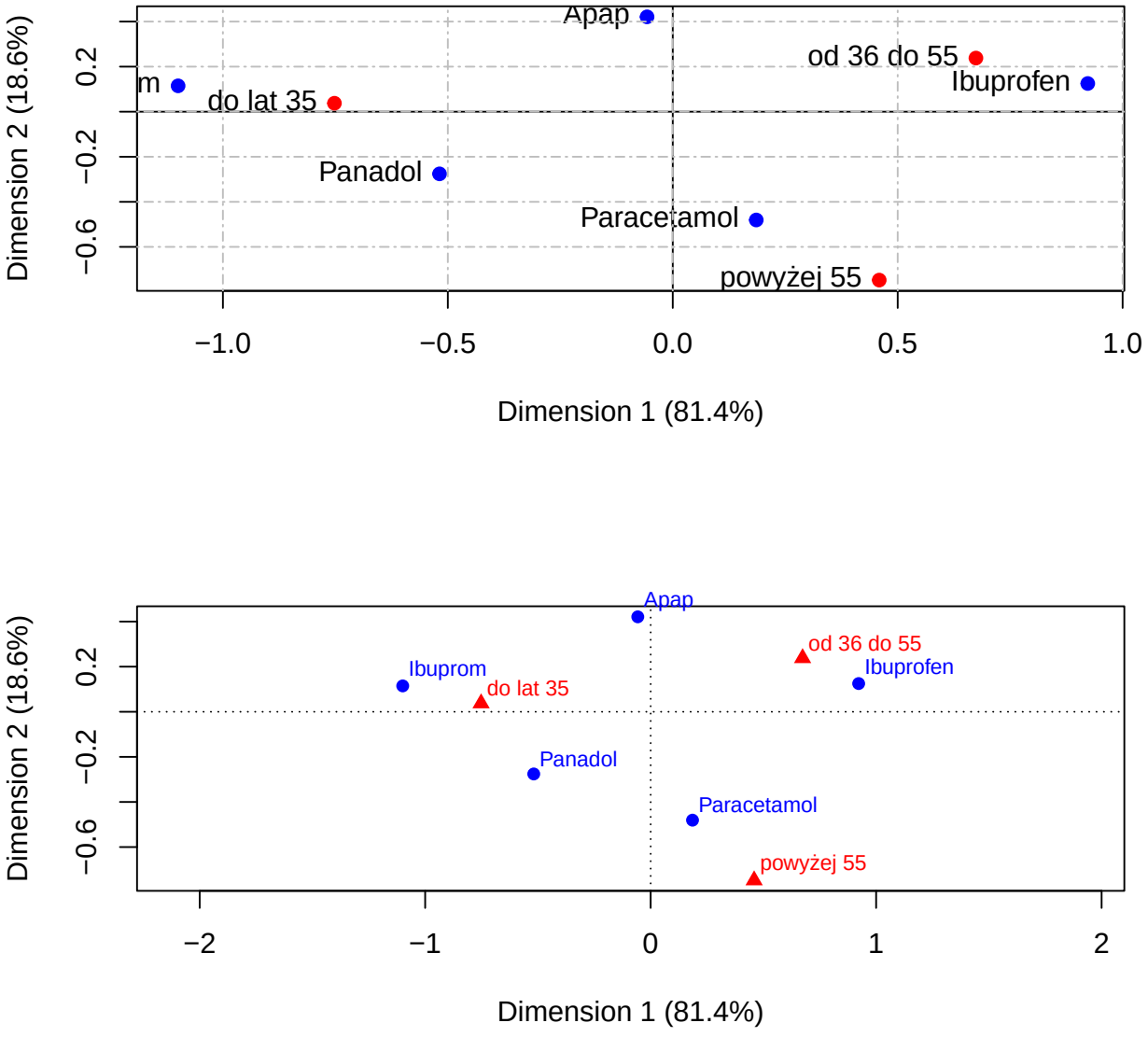
```
## [1] 0.5748397
```

3.2.5. Wnioski

Naszą analizę rozpoczęłem od przeprowadzenia dwóch testów niezależności (Fishera i chi-kwadrat), na poziomie istotności $\alpha = 0.01$. Tak jak w przypadku poprzedniego zadania, p-wartość uzyskana z testu Fishera wynosi dokładnie tyle samo co, dla zmiennych z poprzedniego zadania, więc jeszcze bardziej utwierdza to w przekonaniu, że założenia do przeprowadzenia tego testu (jak występujące zera w odpowiedziach) nie są spełnione. P-wartość z testu chi-kwadrat jest równa praktycznie równa zero, więc ponownie są podstawy do odrzucenia hipotezy o niezależności H_0, więc można założyć, że nasze zmienne nie są niezależne, czyli badanie współzmienności jest uzasadnione.

W tym przypadku wyznaczone współczynniki mają dość spory rozrzut. Współczynnik Goodmana i Kruskala (z wykładu uznany za ważniejszy od innych) dał wartość w przybliżeniu 0.35. Patrząc na pozostałe współczynniki, najmniejszą wartość dał współczynnik T-Czurpowa (≈ 0.45), a największą wartość — współczynnik φ (≈ 0.76). Można uznać, że nasze zmienne charakteryzują się przeciętną albo wysoką współzmiennością.

Następnie przeszedłem do analizy korespondencji. Procenty zmienności sumują się prawie do 100% więc relacje są dobrze przedstawione. Z wykresu można odczytać, że w przypadku bólu osoby powyżej 55 roku życia stosunkowo najczęściej sięgają po „Paracetamol", osoby do lat 35 po „Ibuprom", a osoby między 36, a 55 rokiem życia, po „Ibuprofen", gdyż odpowiednio te punkty na wykresie znajdują się najbliżej siebie.



Rysunek 5. Porównanie wykresu analizy korespondencji z naszej funkcji (na górze) z wykresem funkcji wbudowanej (na dole), dla danych z zadania 2.

Podobnie jak w poprzednim zadaniu współczynniki zmienności i analizę korespondencji wykonałem, korzystając z naszej funkcji oraz funkcji wbudowanych w pakiecie R i uzyskałem takie same wyniki.

Wartość inercji całkowitej jest równa w przybliżeniu 0.57, jest to znacznie większa wartość niż w zadaniu pierwszym (3.1.5), co jest zgodne z intuicją, ponieważ porównując wykres (5) z wykresami (1), (2), (3), (4) widzimy, że wartości na osiach są znacznie większe, więc rozrzut punktów jest większy.

3.3. Zadanie 3

W tym zadaniu należało obliczyć odpowiednią miarę współzmienności zmiennych *Wynagrodzenie* i *Stopień zadowolenia z pracy* oraz przeprowadzić analizę korespondencji.

3.3.1. Dane do analizy

Będę rozważać następujące dane:

	Stopień zadowolenia z pracy				
Wynagrodzenie	b. niezadow.	niezadow.	zadow.	b. zadow.	Suma
< 6000	32	44	60	70	206
$6000 — 15000$	22	38	104	125	289
$15000 — 25000$	13	48	61	113	235
> 25000	3	18	54	96	171
Suma	62	108	319	412	901

Tabela 8. Tablica dwudzielcza rozważanych danych

```
data2 <- matrix(c(32, 44, 60, 70,
                  22, 38, 104, 125,
                  13, 48, 61, 113,
                  3, 18, 54, 96), byrow=T, nrow=4)
dimnames(data2) <- list(c('<6000', '6000 - 15000', '15000 - 25000', '> 25000'),
                                  c('b. niezadow', 'niezadow.', 'zadow.', 'b. zadow.'))
data2
```

```
##               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
```

3.3.2. Analiza niezależności

Przeprowadziłem test (`chisq.test`), którego celem jest zweryfikowanie hipotezy o niezależności wysokości otrzymywanego wynagrodzenia i stopnia zadowolenia z pracy.

Hipotezy dla testu:

- H_0: *Wynagrodzenie* i *Stopień zadowolenia z pracy* są niezależne
- H_1: *Wynagrodzenie* i *Stopień zadowolenia z pracy* nie są niezależne

```
chisq.test(data2)$p.value
```

```
## [1] 4.867831e-08
```

Uzyskałem p-wartość dla testu chi -kwadrat równą praktycznie zero, więc można założyć, że hipoteza zerowa H_0 jest fałszywa, zatem zakładamy, że wynagrodzenie i stopnień zadowolenia z pracy nie są niezależne.

3.3.3. Miara współzmienności

W tym przypadku błędem byłoby wyznaczanie miar współzmienności, analogicznie jak w zadaniach 1. i 2. ((3.1) i (3.2)), ponieważ nasze dane mają charakter uporządkowany.
W przypadku zmiennych uporządkowanych, stosujemy inne miary współzmienności — w tym zadaniu wyznaczymy miarę γ.

```
GoodmanKruskalGamma(data2)
```

```
## [1] 0.218102
```

3.3.4. Analiza korespondencji

Poniżej przedstawiłem macierze i wykresy, które uzyskałem z naszej funkcji, w celu przeprowadzenia analizy korespondencji. Wartości w macierzy zaokrągliłem do czwartego miejsca po przecinku:

$$\mathbf{P:}\begin{bmatrix}0.0355 & 0.0488 & 0.0666 & 0.0777\\ 0.0244 & 0.0422 & 0.1154 & 0.1387\\ 0.0144 & 0.0533 & 0.0677 & 0.1254\\ 0.0033 & 0.02 & 0.0599 & 0.1065\end{bmatrix},\mathbf{A:}\begin{bmatrix}0.1332 & 0.0582 & -0.0158 & -0.0775\\ -0.0032 & -0.0458 & 0.0511 & -0.0134\\ -0.041 & 0.0504 & -0.046 & 0.0248\\ -0.094 & -0.0634 & ,0.0048 & 0.0735\end{bmatrix} \tag{17}$$

$$\mathbf{R:}\begin{bmatrix}0.2286 & 0.3208 & 0.2608 & 0.1898\end{bmatrix},\mathbf{C:}\begin{bmatrix}0.0777 & 0.1643 & 0.3097 & 0.4484\end{bmatrix} \tag{18}$$

$$\mathbf{F:}\begin{bmatrix}-0.3454 & -0.0025 & 0.0218 & 0\\ 0.03 & -0.1178 & -0.0224 & 0\\ 0.0418 & 0.1567 & -0.0181 & 0\\ 0.308 & -0.0133 & 0.0365 & 0\end{bmatrix},\mathbf{G:}\begin{bmatrix}-0.5944 & -0.0923 & 0.0455 & 0\\ -0.205 & 0.1746 & -0.0289 & 0\\ 0.0263 & -0.1222 & -0.0226 & 0\\ 0.1599 & 0.0364 & 0.0183 & 0\end{bmatrix} \tag{19}$$

Wyznaczyłem inercję całkowitą, korzystając ze wzoru (13). Uzyskałem następującą wartość:

```
data2A
```

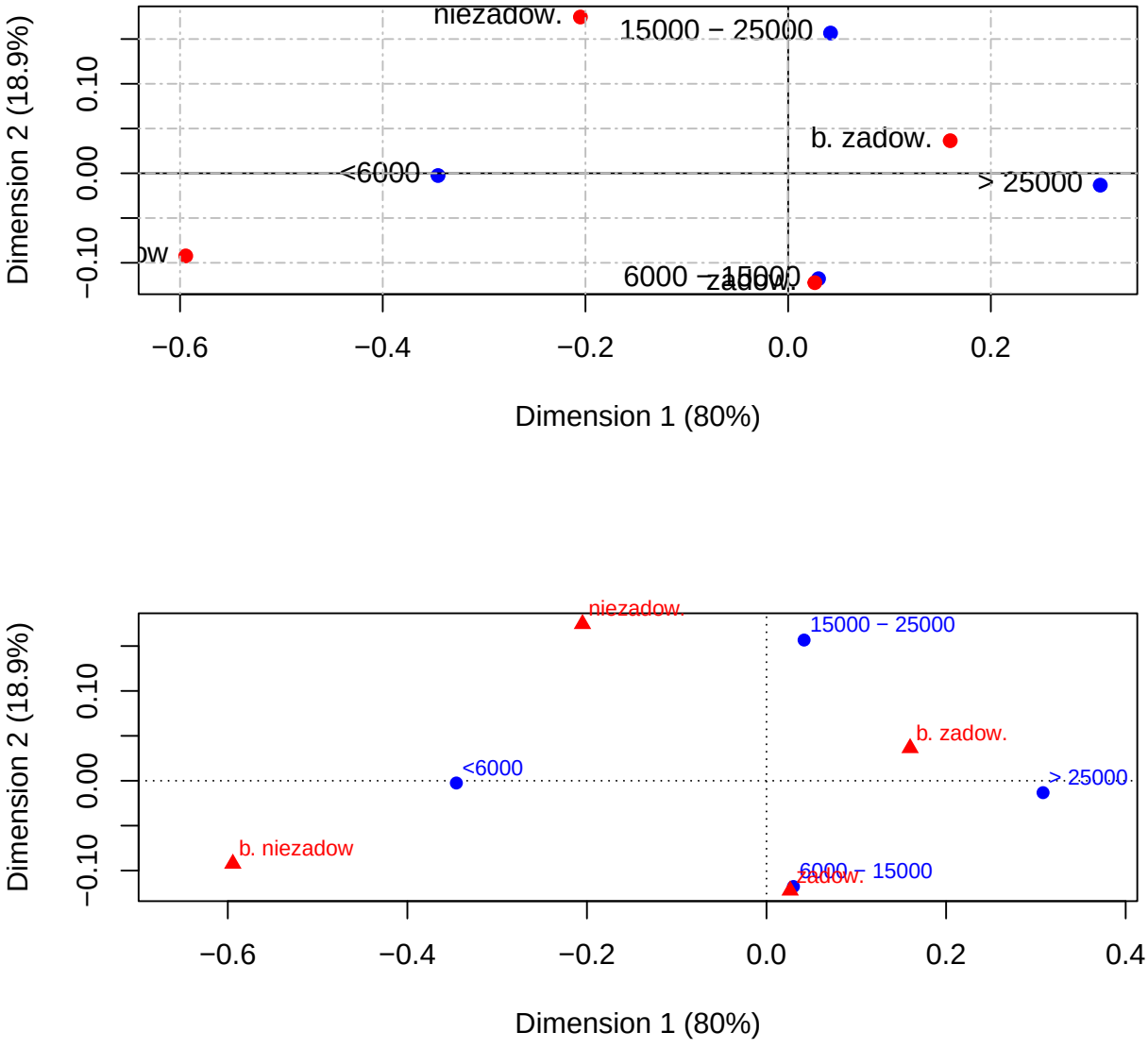
```
##              [,1]        [,2]         [,3]        [,4]
## [1,]  0.133203867  0.05819913 -0.015805310 -0.07753757
## [2,] -0.003183739 -0.04579816  0.051097476 -0.01341809
## [3,] -0.040991673  0.05039544 -0.045963056  0.02475694
## [4,] -0.094008538 -0.06341766  0.004801831  0.07352502
```

```
sum(diag(data2A %*% t(data2A)))
```

```
## [1] 0.05752481
```

3.3.5. Wnioski

Naszą analizę rozpoczęłem od przeprowadzenia testu niezależności chi-kwadrat, na poziomie istotności $\alpha = 0.01$ i uzyskałem p-wartość równą w przybliżeniu $4.9 \cdot 10^{-8}$ więc są podstawy do odrzucenia hipotezy o niezależności H_0 i przyjęcia hipotezy alternatywnej H_1, czyli można



Rysunek 6. Porównanie wykresu analizy korespondencji z naszej funkcji (na górze) z wykresem funkcji wbudowanej (na dole), dla danych z zadania 3.

założyć, że nasze zmienne nie są niezależne. Więc rozsądnym jest badanie współzmienności zmiennych. W tym przypadku mamy zmienne porządkowe (*Wynagrodzenie* i *Stopień zadowolenia z pracy*), więc w tym przypadku zastosowałem inny współczynnik (γ Goodmana i Kruskala) niż w poprzednich zadaniach. Tym razem skorzystałem tylko z funkcji wbudowanej w pakiecie R i uzyskałem wartość w przybliżeniu równą 0.22, przez co można wyciągnąć wniosek, że nasze zmienne charakteryzuje słaba współzmienność.

Na rysunku (6) przedstawiłem wykresy analizy korespondencji. Punkty są dość równomiernie rozrzucone na wykresie, za wyjątkiem punktów reprezentujących osoby ankietowane, zarabiające między 6000, a 15000 i osoby zadowolone z pracy, które praktycznie nachodzą na siebie. Możemy więc wyciągnąć wniosek że dużo osób, których zarobki mieszczą się w przedziale 6000 – 15000 jest zadowolonych z pracy. Osoby ankietowane, zarabiające najwięcej, najczęściej wybierały pozytywne odpowiedzi jeśli chodzi o zadowolenie z pracy. Punkt reprezentujący odpowiedzi „bardzo niezadowolony" lub „niezadowolony" znajduje się najbliżej osób zarabiających mniej niż 6000, więc wśród osób ankietowanych, stosunkowo najczęściej osoby najmniej zarabiające, zaznaczały takie odpowiedzi. Analizę korespondencji wykonałem korzystając z naszej funkcji oraz funkcji wbudowanej i uzyskałem taki sam wykres.

Wartość inercji całkowitej jest równa w przybliżeniu 0.058.