Aleksander Jakóbczyk i Kacper Pasterniak

Sprawozdanie 2

Lista 5, 6 i 7

Zad 1

binom.test jest to funkcja przeprowadzająca test dokładności dla prostej hipotezy zerowej H_0 o prawdopodobieństwu sukcesu w próbie Bernoulliego.

Użycie:

binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95)

gdzie:

- x jest liczbą sukcesów,
- n to liczba prób,
- p to prawdopodobieństwo sukcesu,
- alternative oznacza hipotezę alternatywną H_1,
- conf.level - poziom ufności.

prop.test jest to funkcja sprawdzająca czy prawdopodobieństwa sukcesu w kilku próbkach są takie same, oraz czy są równe pewnym wartością.

Użycie:

prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95, correct = TRUE)

gdzie:

- x jest wektorem z liczbą sukcesów w każdej grupie,
- n to wektor z liczbą prób w każdej grupie,
- p to wektor z prawdopodobieństwami sukcesu w każdej grupie,
- alternative oznacza hipotezę alternatywną H_1,
- conf.level - poziom ufności,
- correct - definiujemy, czy funkcja powinna zastosować poprawkę na ciągłość.

W powyższych testach pojawia się argument *alternative*, oznacza on jaką postać hipotezy alternatywnej H_1 chcemy rozważać:
→ alternative="two.sided", wtedy: $H_1 : \theta_1 \neq \theta_2$,
→ alternative="greater", wtedy: $H_1 : \theta_1 > \theta_2$,
→ alternative="less", wtedy: $H_1 : \theta_1 < \theta_2$,

Zad 2

Załóżmy, że 200 losowo wybranych klientów (w różnym wieku) kilku (losowo wybranych) aptek zapytano, jaki lek przeciwbólowy zwykle stosują. Zebrane dane zawarte są w tablicy 1.

	Wiek ankietowanych			
Lek	do lat 35	od 36 do 55	powyżej 55	Suma
Lbuprom	35	0	0	35
Aap	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. Dane do zad 2. i 3.

a)

Prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap jest mniejsze bądź równe 1/4:

```
binom.test(44, 200, p=1/4, alternative = "g")$p.value

## [1] 0.8562401

prop.test(44, 200, p=1/4, alternative = "g", correct = T)$p.value

## [1] 0.8154462

prop.test(44, 200, p=1/4, alternative = "g", correct = F)$p.value

## [1] 0.8364066
```

Każda z powyższych p-wartości jest większa niż 0,05, zatem w żadnym z powyższych testów nie możemy odrzucić hipotezy, że prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap jest mniejsze bądź równe 1/4.

b)

Prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap jest równe 1/2:

```
binom.test(44, 200,p = 0.5, alternative = "t")$p.value

## [1] 6.837911e-16

prop.test(44, 200, p = 0.5, alternative = "t", correct = T)$p.value

## [1] 4.197514e-15

prop.test(44, 200, p = 0.5, alternative = "t", correct = F)$p.value

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

Widzimy zatem, że p-wartości w każdym teście są mniejsze niż 0,05, zatem odrzucamy hipotezę, że prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Apap jest równe 1/2.

c)

Prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Ibuprom jest większe bądź równe 1/5:

```
binom.test(50, 200,p=1/5, alternative = "l")$p.value

## [1] 0.9655032

prop.test(50, 200, p=1/5, alternative = "l", correct = T)$p.value

## [1] 0.9534609

prop.test(50, 200, p=1/5, alternative = "l", correct = F)$p.value

## [1] 0.9614501
```

Każda z powyższych p-wartości jest większa niż 0,05, zatem w zadanym z powyższych testów nie możemy odrzucić hipotezy, że prawdopodobieństwo, że losowo wybrana osoba z badanej populacji w przypadku bólu zażywa Ibuprom jest większe bądź równe 1/5.

d)

Powtórzyć punkt (a), (b) i (c), ale dla osoby z badanej populacji do lat 35:

```
binom.test(22, 90, p=1/4, alternative = "g")$p.value

## [1] 0.5885826

prop.test(22, 90, p=1/4, alternative = "g", correct = T)$p.value

## [1] 0.5

prop.test(22, 90, p=1/4, alternative = "g", correct = F)$p.value

## [1] 0.5484381
```

Tak samo jak w podpunkcie **a)** nie odrzucamy badanej hipotezy.

```
binom.test(22, 90,p = 0.5, alternative = "t")$p.value

## [1] 1.249258e-06

prop.test(22, 90, p = 0.5, alternative = "t", correct = T)$p.value

## [1] 2.101436e-06

prop.test(22, 90, p = 0.5, alternative = "t", correct = F)$p.value

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

Tak samo jak w podpunkcie **b)** odrzucamy badaną hipotezę.

```
binom.test(0, 90,p=1/5, alternative = "l")$p.value
## [1] 1.897138e-09
prop.test(0, 90, p=1/5, alternative = "l", correct = T)$p.value
## [1] 1.997379e-06
prop.test(0, 90, p=1/5, alternative = "l", correct = F)$p.value
## [1] 1.050718e-06
```

W tym przypadku widzimy że p-wartości w każdym z wykonanych testów jest bliska 0, więc odrzucamy badaną hipotezę.

Zad 3

Na podstawie danych w tablicy 1, korzystając z testu Fishera, na poziomie istotności $\alpha = 0.05$, zweryfikować hipotezę, że prawdopodobieństwo, że osoba do lat 35 zażywa Panadol jest równe prawdopodobieństwu, że osoba od 36 lat do 55 lat zażywa Panadol.

Czy na podstawie uzyskanego wyniku można (na zadanym poziomie istotności) odrzucić hipotezę o niezależności wyboru leku Panadol w leczeniu bólu od wieku, przy uwzględnieniu tylko dwóch grup wiekowych - do lat 35 i od 36 do 55 lat?

fisher.test jest to funkcja przeprowadzająca test dokładności dla hipotezy zerowej H_0 o niezależności kolumn i wierszy w tabeli dwudzielczej.

Użycie:

fisher.test(x, alternative = "two.sided", conf.level = 0.95, simulate.p.value = FALSE)

gdzie:

- x jest dwuwymiarową macierzą,
- alternative oznacza hipotezę alternatywną H_1,
- conf.level - poziom ufności,
- simulate.p.value ustawienie tego paramtru na True sprawi, że metoda sama wysymuluje p-wartości za pomocą metody Monte Carlo,
- ☐ funkcja posiada również inne argumenty, takie jak np. *workspace,hybrid,percent* itp., jednakże nie używamy ich więc objaśnianie jest zbędne.

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

Tabela 2. Tabela do zadania 3.

```r
fisher.test(data2)$p.value
```

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

Jak widzmy p-wartość jest mniejsza od przyjętego poziomu istotności $\alpha = 0.05$, zatem odrzucamy hipotezę o niezależności rozkładów. Zatem badane rozkłady warunkowe nie są jednorodne.

Zad 4

Korzystając z funkcji chisq.test w pakiecie R, na poziomie istotności 0.05, zweryfikować hipotezę o niezależności stopnia zadowolenia z pracy i wynagrodzenia na podstawie danych w tablicy 2. Zwrócić uwagę na stosowaną w tej funkcji poprawkę.

	Zadowolenia				
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
powyzej 25000	7	18	54	92	171
Suma	62	108	5	26	901

Tabela 3. Dane do zadania 4.

```r
chisq.test(data3[1:4,1:4],correct = F)$p.value
```

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

Jak możemy zobaczyć p-wartość jest większa od przyjętego poziomu istotności $\alpha = 0.05$, zatem nie mamy podstaw do odrzucenia hipotezy o niezależności rozkładów warunkowych.

Zad 5

Napisać deklarację funkcji, która dla danych w tablicy dwudzielczej oblicza wartość poziomu krytycznego (p-value) w asymptotycznym teście niezależności opartym na ilorazie wiarogodności. Korzystając z napisanej funkcji, obliczyć tę wartość dla danych z zadania 4.

Wyliczenie p-wartości dla testu opartego na ilorazie wiarogodności oparte jest na statyscyce:

$$G^2 = -2\log(\lambda), \quad \lambda = \prod_{i,j} \left(\frac{n_{i+}n_{+j}}{nn_{ij}} \right)^{n_{ij}},$$
$$\text{p-value} = 1 - F_{\chi^2(R-1)(C-1)}(G^2),$$

gdzie:
n_{i+} - suma liczności w i-tym wierszu,
n_{+j} - suma liczności w j-tej kolumnie,
n_{ij} - liczność w i-tym wierszu i j-tej kolumnie,
n - liczba wszystkich ankietowanych,
R - liczba wierszu testowanej tabeli,
C - liczba kolumn testowanej tabeli.
Powyższa statystyka przy założeniu hipotezy o niezależności, dąży przy $n \to \infty$ do rozkładu χ^2 z $(R-1)(C-1)$ stopniami swobody.

```r
p <- function(dane) {
  y_xo <- rowSums(dane)
  y_ox <- colSums(dane)

  n <- sum(dane)
  r <-nrow(dane)
  c <- ncol(dane)
  x <- 1
  for (i in 1:r) {
    for (j in 1:c) {
      y_ij <- dane[i,j]
      x <- ( ( y_xo[i]*y_ox[j] ) / (y_ij*n) )^y_ij * x
    }
  }
  g_2 <- -2 * log(x)
  return( 1-pchisq(g_2, (c-1)*(r-1)) )
}
p(data3[1:4,1:4])[[1]]
```

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

Jak możemy zobaczyć p-wartość jest większa od przyjętego poziomu istotności $\alpha = 0.05$, zatem nie mamy podstaw do odrzucenia hipotezy o niezależności rozkładów warunkowych. Porównując p-wartości w zadaniu czwarym i piatym których warotści wynosiły odpowiednio 0.2139542 i 0.2112391, widzimy że obie metody w tym przypadku daja wyniki bardzo do siebie zbliżone.

Lista 8 i 9

W kolejnych zadaniach będziemy zajmować się zmiennością danych. Będziemy ją mierzyć za pomocą kilku współczynników, które zdefiniujemy poniżej:
— współczynnik V Cramera:

$$V = \sqrt{\frac{X^2}{n \min\{R-1, C-1\}}}, \tag{1}$$

— współczynnik T-Czuprowa:

$$T = \sqrt{\frac{X^2}{n\sqrt{(R-1)(C-1)}}}, \tag{2}$$

— współczynik φ:

$$\varphi = \sqrt{\frac{X^2}{n}} \tag{3}$$

— współczynnik C Pearsona (coefficient of contingency):

$$C = \sqrt{\frac{X^2}{X^2+n}} \tag{4}$$

— współczynnik $\hat{\gamma}$:

$$\hat{\gamma} = \frac{N_c - N_d}{N_c + N_d} \tag{5}$$

— współczynnik $\hat{\tau}$ Goodmana

$$\tilde{\tau} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}^2}{nn_{i+}} - \sum_{j=1}^{C}\left(\frac{n_{+j}}{n}\right)^2}{1-\sum_{j=1}^{C}\left(\frac{n_{+j}}{n}\right)^2} \tag{6}$$

Powyższa miara okleśla stopień redukcji zmienności zmiennej W_2, przy znajomości zmiennej W_1. Przyjmuje on wartości z zakresu $[0, 1]$. Gdy $\tau = 0$, to wtedy i tylko wtedy zmienne W_1 i W_2 są niezależne. $\tau = 1$, wtedy i tylko wtedy, gdy dla każdego $i \in \{1, ..., R\}$ istnieje $j \in \{1, ..., C\}$ takie, że $p_{j|i} = 1$.

Powyższe miary zdefiniowaliśmy w R:

```
tau <- function(data) {
  p <- data/sum(data)
  pi. <- rowSums(p)
  p.i <- colSums(p)
  R <- nrow(data)
  C <- ncol(data)
  d1 <- 0
  for (i in 1:R) {
    for (j in 1:C) {
      d1 <- p[i, j]^2/pi.[i] + d1
    }
  }
  d2 <- sum(p.i^2)
  return( ( d1 - d2 )/(1 - d2) )}
```

```
wspolczynniki <- function(dane,metoda){
  X <- chisq.test(dane)$statistic[[1]]
  R <- nrow(dane)
  C <- ncol(dane)
  n <- sum(dane)
  if (metoda =="v-cramer"){
    sqrt(X/(n*min(R-1,C-1)))
  }else if (metoda == 't-czuprow'){
    sqrt(X/(n*sqrt((R-1)*(C-1))))
  }else if (metoda=='phi'){
    (X/n)**0.5
  }else if (metoda=='c-pearson'){
   (X/(X+n))**0.5
  }else if (metoda=='tau'){
    tau(dane)
  }else if (metoda=='gamma'){
    a <- 0
    for (i in 1:nrow(dane)-1){
      for (j in 1:ncol(dane)-1){
        a <- a + sum(dane[i,j]) *
            sum(dane[min(i+1,nrow(dane)):nrow(dane),
                     min(j+1,ncol(dane)):ncol(dane)])
      }
    }
    b <-0
    for (i in 1:nrow(dane)-1){
      for (j in ncol(dane):2){
        b <- b + sum(dane[i,j]) *
          sum(dane[min(i+1,nrow(dane)):nrow(dane),
                   max(j-1,1):1])
      }
```

```
}
(a-b)/(a+b)}}
```

Inercja

Inercję całkowitą możma zinterpretować jako miarę rozporoszenia profili w przestrzeni wielowymiarowej. Im większa jest inercja całkowita układu, tym punkty w tej przestrzeni są bardziej rozproszone.

$$\lambda = \sum_{i=1}^{K} \gamma_i^2 \tag{7}$$

gdzie: λ_i to wartości osobliwe macierzy A, które jednocześnie znajdują się w diagonalnej macierzy Γ, możemy również wyznaczyć współczynik inercji cześciowej:

$$\lambda_c = \frac{\sum_{i=1}^{2} \gamma_i^2}{\lambda} \tag{8}$$

Zad. 1

Tabele z przykładu 1 wykład 7:

```
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)
m2a=matrix(c(505, 202, 19, 136,
             240,  77, 14,  88,
             181,  63,  8, 105,
             512, 159, 21, 294),nrow=4,byrow=TRUE)
m1b=matrix(c(888, 369, 50, 457,
             263,  95, 10,  99,
             208,  29,  2,  44,
              79,   9,  0,  23),nrow=4,byrow=TRUE)
m1c=matrix(c(888, 369, 50, 457,
             263,  95, 10,  99,
             287,  38,  2,  67),nrow=3,byrow=TRUE)
```

	τ	V	T	φ	C
m1a	0.01712	0.10292	0.09578	0.17827	0.17550
m1b	0.01508	0.09627	0.09627	0.16675	0.16448
m1c	0.01490	0.11682	0.10556	0.16521	0.16300
m2a	0.01013	0.09116	0.09116	0.15789	0.15596

Tabela 4. Tablica wartości współczynników dla poszczegulnych macieży dla adania 1.

Jak możemy zobaczyć otrzymana wartość współczynnika τ jest bliska zeru, świadczy to o bardzo małej współzmienności badanych zmiennych losowych.

```r
inercja <- function(data) {
  p <- data/sum(data)
  r <- rowSums(p)
  c <- colSums(p)
  R <- nrow(data)
  C <- ncol(data)
  P<- matrix(p,R,C)
  D_r <- diag(r)
  D_c <- diag(c)
  A <- sqrt( inv(D_r) ) %*% ( P - r %*% t(c) ) %*% sqrt( inv(D_c) )
  B <- diag(A %*% t(A))
  g <- diag(svd(A)$d)
  return(c( sum( B ), (g[1,1]^2+g[2,2]^2)/sum(g^2) ) )
}
```

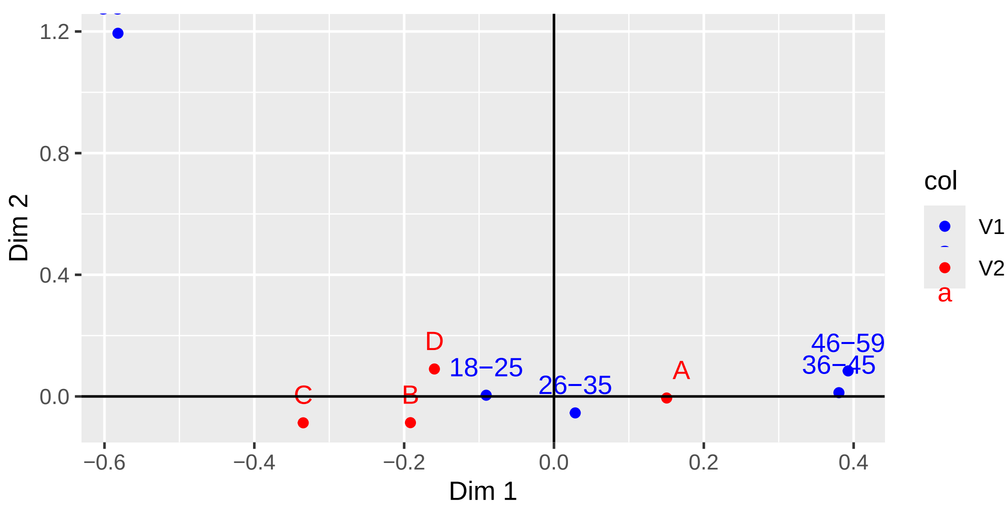
	λ	inercja cz.
m1a	0.03178	0.99631
m1b	0.02780	0.99648
m1c	0.02730	1.00000
m2a	0.02493	0.99586

Tabela 5. tablica wartości współczynników inercji macierzy dla zadania 1.

Dla badanych tabel przeprowadzona analiza korespodencji wykazała, że dwa wymiary wystarczą, aby objaśnić ponad 99.5% zmienności, a w tabeli m1c nawet 100%. W każdym przypadku incercja całkowita jest bardzo mała, co świadczy o niewielkim rozproszeniu punktów w przestrzeni wielowymiarowej.

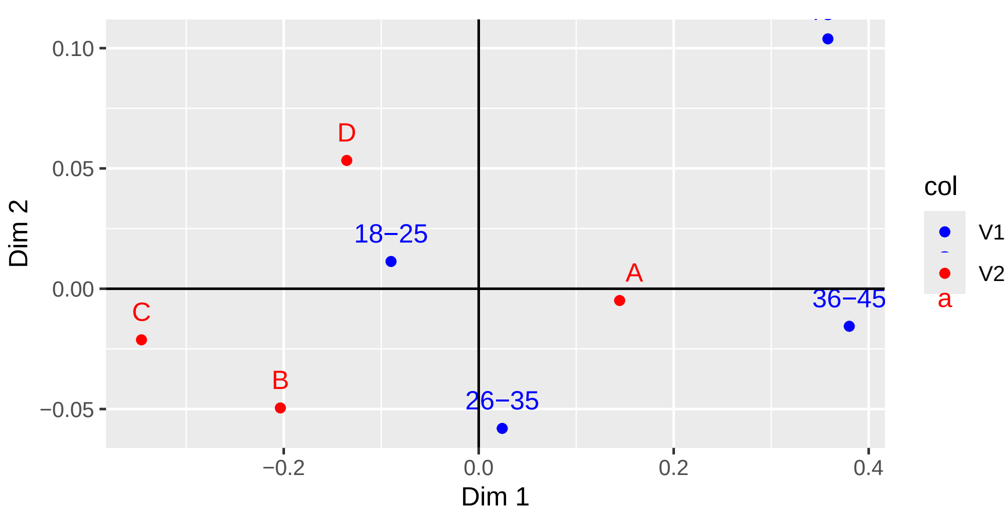
```r
analiza <- function(data) {
  p <- data/sum(data)
  r <- rowSums(p)
  c <- colSums(p)
  R <- nrow(data)
  C <- ncol(data)
  P<- matrix(p,R,C)
  D_r <- diag(r)
  D_c <- diag(c)
  A <- sqrt( inv(D_r) ) %*% ( P - r %*% t(c) ) %*% sqrt( inv(D_c) )
  s <- svd(A)
  U <- s$u
  D <- diag(s$d)
  V <- s$v
  F <- sqrt( inv(D_r) ) %*% U %*% D
  G <- sqrt( inv(D_c) ) %*% V %*% D
  F_df <- data.frame(
    x = c(F[,1],G[,1]),
```

```
    y = c(F[,2],G[,2]),
    name = c(rownames(p),colnames(p)),
    col = c(rep("V1",R),rep("V2",C))
  )
  return(F_df)
}
```



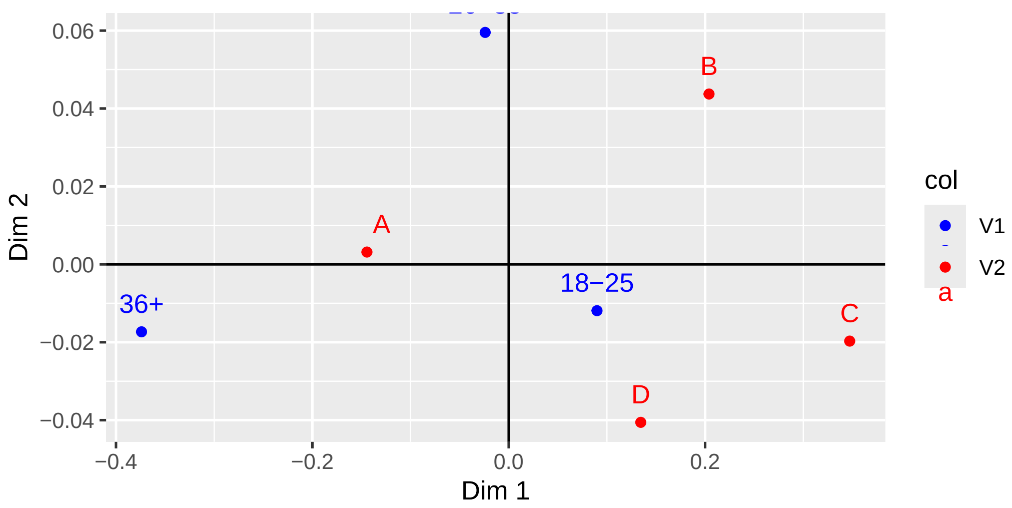
Rysunek 1. Analiza korespondencji dla tabeli m1a.

Na wykresie zauważamy, że nie istnieją silne powiązania między zmiennymi, wszystkie skupione są blisko początku układu współrzędnych (punktu (0,0)). Kategoria "60+" cechuje się zdecydowanie największa rozróżnialnościa, na co wskazuję duża odległość od wcześniej wspomnianego punktu.



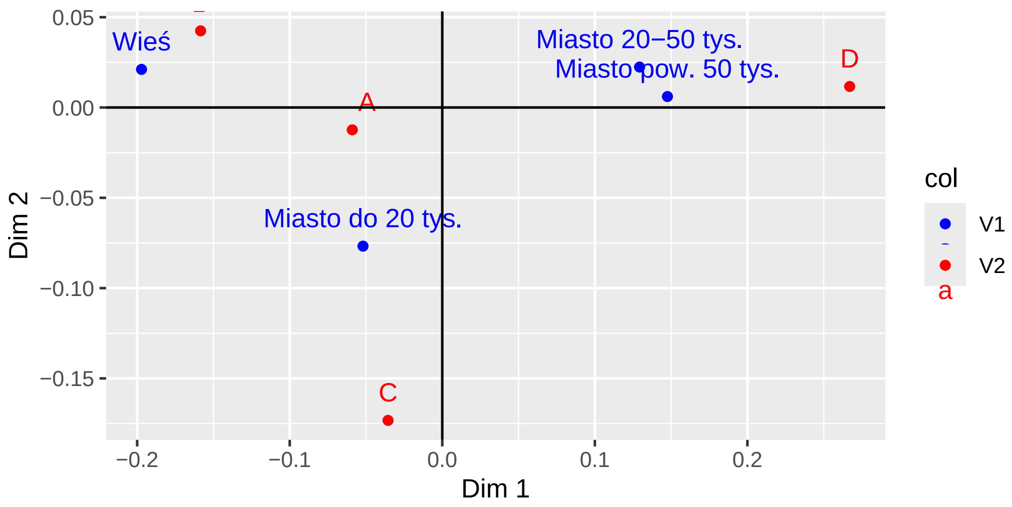
Rysunek 2. Analiza korespondencji dla tabeli za m1b.

Połączenie kategori "60+" z kategorią "46-59" nie zmienia znacząco wartości ponktów na wykresie, jest to spowodowane tym, że w kategori "60+" znajduje sie tylko 8 osób.



Rysunek 3. Analiza korespondencji dla tabeli za m1c.

Połączenie kategorii "46+" z kategorią "36-45" wpłynęło na kształt wykresu, teraz osoby w wieku "18-25" wykazują największe powiązanie ze "segreguje śmieci, ponieważ każdy tak robi" i nie są skłonni do niesegregowania śmieci. Natomiast kategorie "36+" i "26-35" są najbardziej oddalone od początku układu współrzędnych co wskazuje, że są najbardziej rozróżnialne.



Rysunek 4. Analiza korespondencji dla tabeli za m2a.

Z wykresu możemy wywnioskować, że osoby z miast liczących ponad 50 tysięcy mieszkańców i od 20 do 50 tysięcy mieszkańców są najbardziej związani z odpowiedzią "nie segreguje śmieci". Spora korespondencja występuje między mieszkańcami wsi a odpowiedzą "segreguje smeici ponieważ tak nakazuje prawo", a mieszkańcy miast do 20 tysięcy mają najsilniejsze powiązanie z odpowiedzią "segreguje śmieci, ponieważ wszyscy tak robią"

Zad 2

Załóżmy, że 200 klientów (w różnym wieku) kilku aptek zapytano, jaki lek przeciwbólowy zwykle stosują. Zebrane dane zawarte są w tablicy 2. Na podstawie tych danych, obliczyć odpowiednie miary współzmienności oraz przeprowadzić analizę korespondencji, tzn. obliczyć wartości odpowiednich macierzy, współrzędnych punktów oraz utworzyć odpowiednie wykresy.

	Wiek ankietowanych		
Lek	do lat 35	od 36 do 55	powyżej 55
Ibuprom	35	0	0
Aap	22	22	0
Paracetamol	15	15	15
Ibuprofen	0	40	10
Panadol	18	3	5

Tabela 6. Dane do listy 8 i 9, zadanie 2.

	τ	V	T	φ	C
1	0.34772	0.53612	0.45082	0.75818	0.60416

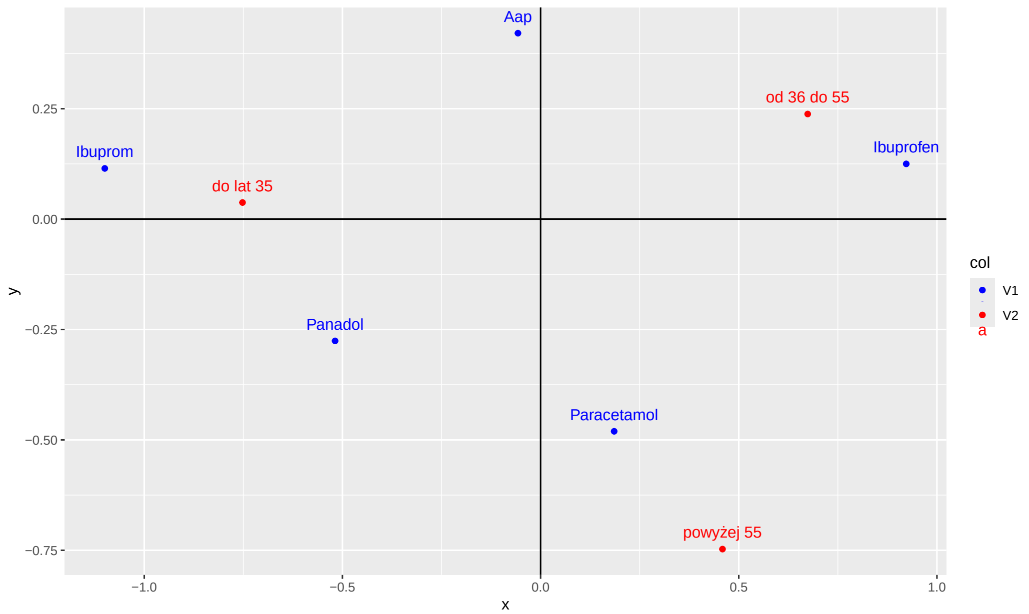
Tabela 7. Tablica wartości współczynników dla zadania 2.

Tabela w tym zadaniu nie jest rozmiaru 2x2, więc powinniśmy rozpatrzyć tylko współczynnik τ. Jak możemy zobaczyć otrzymana wartość tego współczynnika wskazuję na to, że zmienna *Lek* ma wpływ na redukcję zmienności zmiennej *Wiek*.

	λ	inercja cz.
Zad 2	0.57484	1.00000

Tabela 8. tablica wartości współczynników inercji macierzy dla zadania 2.

Z przeprowadzonej analizy korespodencji wynika, że incercja całkowita wynosi 0.57484, a w dwóch wymiarach objaśnione jest aż 100% zmienności.



Rysunek 5. Analiza korespondencji dla Tabeli 1.

Na wykresie widzimy, że silne powiązanie występuje między zmiennymi "Ibuprom" oraz "do lat 35", zmienna "Paracetamol" także jest silnie powiązana, ale ze zmienną "powyżej 55". Natomiast lek "Ibuprofen" wykazuje najsilniejszą korespodencję z grupą wiekową "od 36 do 55".

Zad 3

Na podstawie danych zawartych w tablicy 1, obliczyć (odpowiednią) miarę współzmienności zmiennych Wynagrodzenie i Stopień zadowolenia z pracy (dane te są trochę inne niż te rozpatrywane na poprzednij liście). Następnie, przeprowadzić analizę korespondencji, tzn. obliczyć wartości odpowiednich macierzy, współrzędnych punktów oraz utworzyć odpowiednie wykresy.

	Stopień zadowolenia z pracy			
Wynagrodzenie	b. niezadow.	niezadow.	zadow.	b. zadow.
do 6000	32	44	60	70
6000-15000	22	38	104	125
15000-25000	13	48	61	113
powyzej 25000	3	18	54	96

Tabela 9. Dane do listy 8 i 9, zadanie 3.

	τ	V	T	φ	C	$\hat{\gamma}$
1	0.01671	0.13847	0.13847	0.23984	0.60416	0.21810

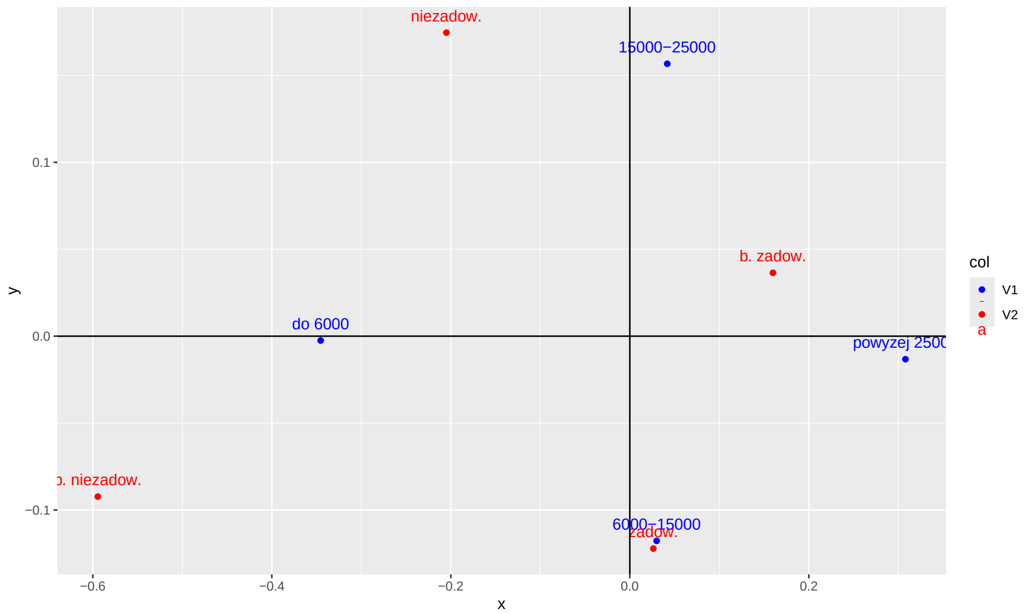
Tabela 10. Tablica wartości współczynników dla zadania 3.

Ze względu na to ze badana tabela zawiera dwie zmienne kategoryczne i uporządkowane, możemy w tym przypadku posłużyć się dodatkowym współczynnikiem $\hat{\gamma}$.

	λ	inercja cz.
Zad 3	0.05752	0.98946

Tabela 11. Tablica wartości współczynników inercji macierzy dla zadania 3.

Inercja całkowita wynosi 0.05752, a częściowa aż 0.98946, co oznacza, że aż 98.95% zmienności jest objaśnione w dwóch wymiarach.



Rysunek 6. Analiza korespondencji dla Tabeli 3.

Wartość $\hat{\gamma}$ wskazuję, że między zmiennymi "Wynagrodzenie" oraz "Stopień zadowolenia z pracy" istnieje współzmienność. Z powyższego wykresu korespodencji możemy wywnioskować, że najsłabiej rozróżnialnymi zmiennymi są kategorie "b.zadow" oraz "zadow" i "6000-15000", wskazuję na to mała odległość od punktu *(0,0)* na wykresie. Najdalej, od wcześniej wspomnianego puntku, leży natomiast odpowiedź "b.niezadow", co wskazuje, że to ta kategoria jest najbardziej rozróżnialna. Silne powiązanie występuje między odpowiedziami "b.niezadow" i "do 6000".