Sprawozdanie 2

Analiza danych ankietowych

Aleksandra Grzeszczuk
album 255707

Jacek Wszoła
album 255718

5 stycznia 2022

Spis treści

1 Lista 1

1.1 Zadanie 1

Dla rozkładu wielomianowego $\mathcal{M}_3(n, \mathbf{p})$ z parametrem $\mathbf{p} = (p_1, p_2, p_3)$ porównamy dwa estymatory i na podstawie symulacji sprawdzimy, który z nich lepiej estymuje nieznany parametr. Dla $n = 1, 2, 3$ mamy estymatory:

$$\widehat{p}_i = \frac{X_i}{n} \quad \text{oraz} \quad \widetilde{p}_i = \frac{X_i + 1/2}{n + 3/2}.$$

Jako wyznacznik przyjmiemy średnią wartość różnicy kwadratów. Poniżej prezentujemy funkcję, której użyjemy do przeprowadzenia symulacji dla różnych wartości $\mathbf{p}$.

```
estymatory <- function(p){
  est.hat <- c()
  est.tilde <- c()
  for(n in c(50, 100, 1000)){
    X <- rmultinom(1000, size = n, prob = p)
    p.hat <- X/n
    p.tilde <- (X+1/2)/(n+3/2)
    P <- matrix(rep(p, times = 1000), nrow = 3, byrow = FALSE)
    est.hat <- append(est.hat, 1/1000 * sum((P-p.hat)^2))
    est.tilde <- append(est.tilde, 1/1000 * sum((P-p.tilde)^2))
  }
  zestawienie <- as.data.frame(matrix(c(est.hat, est.tilde),
                                      nrow = 2, byrow = TRUE))
  colnames(zestawienie) <- c("$n=50$", "$n=100$", "$n=1000$")
  rownames(zestawienie) <- c("dla $\\mathbf{\\widehat{p}(X)}$",
                             "dla $\\mathbf{\\widetilde{p}(X)}$")
  tab <- xtable(zestawienie, row.names = FALSE, digits = 10,
         caption = "Kryterium porównawcze dla estymatorów")
  print(tab, type = "latex", table.placement = "H",
        sanitize.text.function=function(x){x})
}
```

(a) $\mathbf{p} = (1/3, 1/3, 1/3)$.

	$n = 50$	$n = 100$	$n = 1000$
dla $\widehat{\mathbf{p}}(\mathbf{X})$	0.0126034667	0.0068186667	0.0006686887
dla $\widetilde{\mathbf{p}}(\mathbf{X})$	0.0118799761	0.0066186189	0.0006666871

Tabela 1: Kryterium porównawcze dla estymatorów

Przypomnijmy, że im średnia wartość różnicy kwadratów mniejsza, tym dokładniejsze przybliżenie. W tym przypadku oba estymatory dają podobne wyniki, z delikatną przewagą $\widetilde{\mathbf{p}}$, jest to jednak znikoma różnica. Rosnącą dokładność wraz ze wzrostem wielokości próby – dla $n = 50$ mamy błąd rzędu 0.01, zaś dla $n = 1000$ jest to już rząd 0.0001.

(b) $\mathbf{p} = (1/10, 4/5, 1/10)$.

	$n = 50$	$n = 100$	$n = 1000$
dla $\widehat{\mathbf{p}}(\mathbf{X})$	0.0065256000	0.0034448000	0.0003486420
dla $\widetilde{\mathbf{p}}(\mathbf{X})$	0.0064843435	0.0034623699	0.0003486076

Tabela 2: Kryterium porównawcze dla estymatorów

Tendencja jest podobna do punktu (a). Oba estymatory zachowują tę samą dokładność z małą przewagą $\widetilde{\mathbf{p}}$. Jednak w tym przypadku obserwujemy znacznie mniejszą wartość błędu, który jest o około połowę mniejszy dla każdej z badanych wielkości prób.

(c) $\mathbf{p} = (2/5, 1/5, 2/5)$.

	$n = 50$	$n = 100$	$n = 1000$
dla $\widehat{\mathbf{p}}(\mathbf{X})$	0.0125608000	0.0064628000	0.0006303560
dla $\widetilde{\mathbf{p}}(\mathbf{X})$	0.0118662268	0.0062721444	0.0006287689

Tabela 3: Kryterium porównawcze dla estymatorów

Zauważamy podobną sytuację jak w punkcie (a). Również rzędy przybliżeń zgadzają się z pierwszym przypadkiem.

Biorąc pod uwagę powyższe przypadki, możemy sformułować hipotezę, że im większa amplituda prawdopodobieństw w wektorze $\mathbf{p}$, tym mniejsza precyzja estymacji dla obu estymatorów. W każdym badanym przypadku wielkość próby jest proporcjonalna do dokładności. Co więcej, za każdym razem lepsze wyniki osiągał estymator $\widetilde{\mathbf{p}}$, więc to jego powinniśmy wybrać.

1.2 Zadanie 2

Zajmiemy się teraz testowaniem hipotez dotyczących rozkładu wielomianowego $\mathcal{M}_k(n, \mathbf{p})$. Przyjrzymy się problemom testowania hipotez typu

$$H_0 : \mathbf{p} = \mathbf{p}_0 \quad \text{przeciwko} \quad \mathbf{p} \neq \mathbf{p}_0.$$

W tym celu napiszemy funkcje, które będą zwracać p-value dla trzech możliwych do wyboru przez użytkownika testów: χ^2 Pearsona, χ^2 największej wiarygodności oraz Walda.

```
hip.test <- function(X, p0, type = "Pearson"){
  k <- length(p0)
  n <- sum(X)
  if(type == "Pearson"){
    chi2 <- sum((X-n*p0)^2/(n*p0))
    return(1-pchisq(chi2, df = k-1))
  }
  if(type == "ML"){
    G2 <- 2*sum(X*log(X/(n*p0)))
    return(1-pchisq(G2, df = k-1))
  }
```

```
  if(type == "Wald"){
    W <- sum((X-n*p0)^2/X)
    return(1-pchisq(W, df = k-1))
  }
}
```

Aby sprawdzić poprawność działania naszej funkcji, możemy posłużyć się danymi Darwina. Ustalił on, że dla wektora obserwacji $X = (315, 108, 101, 32)$ wektor prawdopodobieństw $\mathbf{p}$ określający rozkład, z którego zaobserwowano dane, powinien wynosić $\mathbf{p} = (9/16, 3/16, 3/16, 1/16)$. I właśnie takie $\mathbf{p}_0$ weźmy.

```
X <- c(315, 108, 101, 32)
p0 <- c(9/16, 3/16, 3/16, 1/16)
hip.test(X, p0, type = "Pearson")

## [1] 0.9254259

hip.test(X, p0, type = "ML")

## [1] 0.9242519

hip.test(X, p0, type = "Wald")

## [1] 0.9216972
```

Rzeczywiście, dla wszystkich przypadków p-value wyszło bliskie jedynki, zatem nie mamy podstaw do odrzucenia hipotezy.

1.3 Zadanie 3

W pewnej bardzo dużej korporacji zostały przeprowadzone badania zadowolenia z pracy. Anonimowo zapytano 901 pracowników o ich stopień zadowolenia z pracy. Wyniki były następujące:

	STOPIEŃ ZADOWOLENIA	LICZBA OSÓB
1	bardzo niezadowoleni	62
2	niezadowoleni	108
3	zadowoleni	319
4	bardzo zadowoleni	412

Tabela 4: Stopień zadowolenia z pracy 901 osób

Na podstawie powyższych danych oszacujemy przedziałowo prawdopodobieństwo stopnia zadowolenia na przyjętym poziomie ufności $\alpha = 0.95$. Wykorzystamy do tego gotową funkcję `MultinomCI` dostępną w pakiecie `DescTools`.

```
##             est     lwr.ci    upr.ci
## [1,] 0.06881243 0.03440622 0.1039800
## [2,] 0.11986681 0.08546060 0.1550343
## [3,] 0.35405105 0.31964484 0.3892186
## [4,] 0.45726970 0.42286349 0.4924372
```

Analizując powyższą tabelkę widzimy, że na przyjętym poziomie ufności $\alpha = 0.95$ przedziały ufności wynoszą kolejno:

- Ludzie bardzo niezadowoleni z pracy: 4% – 10% (6%)
- Ludzie niezadowoleni z pracy: 9% – 16% (7%)
- Ludzie zadowoleni z pracy: 32% – 39% (7%)
- Ludzie bardzo zadowoleni z pracy: 42% – 50% (8%)

W każdym z wyznaczonych przedziałów różnica między górnym a dolnym oszacowaniem wynosi około 7%, czyli jest to dość wąski przedział, co oznacza, że nasze dopasowanie jest pewne.

1.4 Zadanie 4

W poprzednim zadaniu oprócz stopnia zadowolenia z pracy pytano również o wysokość ich wynagrodzenia. W grupie 108 osób niezadowolonych z pracy wyniki były następujące:

	WYNAGRODZENIE	LICZBA OSÓB
1	... - 6000$	24
2	6000$ - 15000$	38
3	15000$ - 25000$	28
4	25000$ - ...	18

Tabela 5: Wynagrodzenie badanych 108 osób niezadowolonych z pracy

Na podstawie powyższych danych, na poziomie istotności $\alpha = 0.05$, zweryfikujemy hipotezę, że w grupie pracowników niezadowolonych z pracy, rozkład zarobków w powyższych czterech kategoriach jest równomierny. Wykorzystamy do tego wyżej napisaną funkcję `hip.test`.

```
hip.test(c(24,38,28,18),c(1/4, 1/4, 1/4, 1/4), "Pearson")

## [1] 0.04917479

hip.test(c(24,38,28,18),c(1/4, 1/4, 1/4, 1/4), "ML")

## [1] 0.05125939

hip.test(c(24,38,28,18),c(1/4, 1/4, 1/4, 1/4), "Wald")

## [1] 0.0440901
```

Widzimy, że wyznaczone wartości są prawie identyczne - bardzo małe, wahające się w okolicach 0.05. Ostatecznie stwierdzamy, że w grupie osób niezadowolonych z pracy rozkład zarobków jest równomierny. To znaczy, że jest jednakowe prawdopodobieństwo, że pracownik niezadowolony z pracy zarabia mniej niż 6000$, od 6000$ do 15000$, powyżej 15000$, ale poniżej 25000$ oraz 25000$ lub więcej.

2 Lista 2

2.1 Zadanie 1

W tym zadaniu na podstawie obserwacji startu promu kosmicznego, ściślej: uszkodzeń pierścieni oraz temperatury otoczenia, na poziomie istotności $\alpha = 0.05$ zweryfikujemy hipotezę dotycząca niezależności obu zmiennych. Zaobserwowane zdyskretyzowane wartości przedstawiamy w poniższej tabeli.

	$\leq 65°F$	$> 65°F$	Σ_{uszk}
Nie	0	17	17
Tak	4	3	7
Σ_T	4	20	24

Tabela 6: Obserwacje startu promu kosmicznego

Ponieważ mamy do czynienia z tabelą 2×2, więc do weryfikacji hipotezy o niezależności zmiennych użyjemy testu Fishera. Jest on szczególnym przypadkiem testu Freemana-Haltona. Do zaimplementowania tej metody użyjemy funkcji `fisher.test()`. Otrzymujemy p-value równe około $0.003 < \alpha$, więc hipotezę o niezależności zmiennych odrzucimy z prawdopodobieństwem 1.

```
M <- matrix(c(0, 17, 4, 3), byrow = TRUE, nrow = 2)
fisher.test(M)$p.value

## [1] 0.003293808
```

2.2 Zadanie 2

Jak już wiemy, dane w pliku `Reakcja.csv` zawierają informację o reakcji na lek (zmienna `Reakcja` na poziomie 0, gdy nie nastąpiła poprawa i na poziomie 1, gdy nastąpiła poprawa) w różnych dawkach - zmienna `Dawka`, dwóch firm farmaceutycznych pacjentów leczonych w domu (0) bądź w szpitalu (1).

Na podstawie uzyskanych danych znajdziemy odpowiedz, na pytanie, czy skuteczność leczenia (zmienna `Reakcja`) jest niezależna od wielkości dawki (zmienna `Dawka`). Wykorzystamy w tym celu test chi-kwadrat Pearsona.

```
RD <- table(dane$Reakcja, dane$Dawka)
chisq.test(RD, simulate.p.value = TRUE)

##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  RD
## X-squared = 20.691, df = NA, p-value = 0.0009995
```

Wartość poziomu krytycznego wynosi w tym teście:

```
## [1] 0.00149925
```

Jest ona znikomo mała, zatem odrzucamy hipotezę o niezależnośći skuteczności leczenia od wielkości dawki.

Dalej zbadamy niezależność skuteczności leczenia a rodzaju leku.

```
RR <- table(dane$Reakcja, dane$Rodzaj)
chisq.test(RR)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  RR
## X-squared = 0.41073, df = 1, p-value = 0.5216
```

W tym przypadku wartość poziomu krytycznego wynosi

```
## [1] 0.5215991
```

Co jest wyższe od poziomu istotności $\alpha = 0.05$, zatem przyjmujemy hipotezę zerową o niezależności skuteczności leczenia a rodzaju leku. To oznacza, że obydwa leki działają poprawnie i nie ma między nimi większych dysproporcji.

Ostatecznie zbadamy niezależności skuteczności leczenia od miejsca leczenia.

```
RM <- table(dane$Reakcja, dane$Miejsce)
chisq.test(RM)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  RM
## X-squared = 14.786, df = 1, p-value = 0.0001204
```

Wartość poziomu krytycznego ponownie jest bardzo mała i wynosi zaledwie

```
## [1] 0.0001204076
```

Zatem odrzucamy hipotezę zerową o niezależności skuteczności leczenia od miejsca leczenia.

Podsumowując powyższe zadanie, otrzymane wyniki wydają się bardzo rzeczywiste. Na skuteczność leczenia faktycznie ma wpływ wielkość dawki - czy przyjmiemy jej więcej, czy mniej a także miejsce leczenia. Oczywistym jest, że w przebywając w szpitalu w większości przypadków szybciej wyzdrowiejemy - zajmie się tam nami wyspecjalizowana kadra, zawsze otrzymamy potrzebną, profesjonalną pomoc. Ponadto na skuteczność leczenia nie ma wpływu rodzaj leku - to dobrze, ponieważ obydwa leki muszą bazować na tych samych składnikach i obydwa dają podobne rezulataty.

W zadaniu skorzystałam z testu chi-kwadrat Pearsona, ponieważ jest on najczęściej używanym w praktyce testem. Można go wykorzystać do badania zgodności zarówno cech mierzalnych, jak i niemierzalnych. Ten test wykonuje statystykę testową:

$$X^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(n_{ij} - (n_{i+}n_{+j})/n)^2}{(n_{i+}n_{+j})/n}$$

Dodatkowo, w celu porównania wyników wyznaczę wartości krytyczne korzystając z testu Fishera.

```
fisher.test(RD)$p.value #REAKCJA/DAWKA

## [1] 0.000299323

fisher.test(RR)$p.value #REAKCJA/RODZAJ

## [1] 0.521833

fisher.test(RM)$p.value #REAKCJA/MIEJSCE

## [1] 9.773288e-05
```

Otrzymane wyniki dają nam takie same wnioski jak te opisane powyżej.

2.3 Zadanie 3

Ponownie zajmiemy się danymi przeprowadzonymi w pewnej bardzo dużej korporacji. Tym razem znamy już wynagrodzenie wszystkich osób zadowolonych bądź też nie ze swojej pracy. Korzystając z funkcji `chisq.test` na poziomie istotności $\alpha = 0.05$ zweryfikujemy hipotezę o niezależności stopnia zadowolenia z pracy i wynagrodzenia na podstawie danych zawartych w poniższej tabeli:

	WYNAGRODZENIE	BN	N	Z	BZ
1	... - 6000	20	24	80	82
2	6000 - 15000	22	38	104	125
3	15000 - 25000	13	28	81	113
4	25000 - ...	7	18	54	92

Tabela 7: Wynagrodzenie badanych 901 osób i ich stopień zadowolenia

Gdzie BN - oznacza osoby bardzo niezadowolone z pracy, N - osoby niezadowolone, Z- osoby zadowolone oraz BZ - osoby bardzo zadowolone.

```
m <- matrix(c(20, 24, 80, 82, 22, 38, 104, 125, 13, 28, 81, 113, 7, 18, 54, 92),
            nrow = 4, byrow = TRUE)
chisq.test(m)

##
##  Pearson's Chi-squared test
```

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

Korzystając z testu chi-kwadrat Pearsona bez poprawki, na poziomie istotności $\alpha = 0.05$ nie mamy podstaw do odrzucenia hipotezy zerowej, więc przyjmujemy, że dane są niezależne. Wartość poziomu krytycznego w tym teście wynosi 0.214. Obliczymy teraz wartosc testu chi-kwadrat Pearsona z naniesioną poprawką:

```
chisq.test(m, simulate.p.value = TRUE)

##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  m
## X-squared = 11.989, df = NA, p-value = 0.2104
```

Poziom krytczyny wynosi 0.2154 co nadal jest powyżej poziomu istotności $\alpha = 0.05$. Zatem ponownie nie odrzucamy hipotezy zerowej.

2.4 Zadanie 4

Stworzymy funkcję, która dla danych w tablicy dwudzielczej oblicza wartość p-value w teście niezależności opartym na ilorazie wiarogodności. Wiadomo, że statystyka testowa to

$$G^2 = -\log\lambda, \text{ gdzie } \lambda = \prod_{i,j}\left(\frac{n_{i+}n_{j+}}{nn_{ij}}\right)^{n_{ij}}$$

i dąży ona według rozkładu do rozkładu $\chi^2_{(R-1)(C-1)}$, gdzie R, C to odpowiednio liczba wierszy i kolumn tabeli, a n jest liczbą ankietowanych, tzn. $n = \sum_j n_{j+} = \sum_i n_{i+} = \sum_{i,j} n_{ij}$. Zatem wartość p-value obliczymy ze wzoru $p = 1 - F_{\chi^2_{(R-1)(C-1)}}(G^2(x))$.

```
test.iw <- function(X){
  n <- sum(X)
  C <- ncol(X)
  R <- nrow(X)
  ni <- rowSums(X)
  nj <- colSums(X)
  lambda <- 1
  for(i in 1:R){
    for(j in 1:C){
      lambda <- lambda * ((ni[i]*nj[j])/(n*X[i,j]))^(X[i,j])
    }
  }
  G2 <- -2*log(prod(lambda))
  p <- 1 - pchisq(G2, (R-1)*(C-1))
```

```
  return(p)
}
```

Poprawność funkcji sprawdzimy dla danych z zadania 2.1. Obserwujemy, że wartość funkcji (p-value) dla tych danych jest bliska zeru, a więc hipotezę z prawdopodobieństwem 1 należy odrzucić. Podobny wynik otrzymaliśmy przy użyciu testu Fishera, zatem możemy przypuszczać, że nasza funkcja działa poprawnie.

```
X <- matrix(c(0, 17, 4, 3), byrow = T, nrow = 2)
test.iw(X)

## [1] 0.0005134357
```

3 Lista 3

3.1 Zadanie 1

Ponownie skorzystamy z danych z pliku `reakcja.csv`. Tym razem skupimy się na zmiennych reakcja, dawka, miejsce; zbadamy ich niezależność i, o ile hipotezę o niezależności będziemy mogli odrzucić, obliczymy odpowiednie miary współzmienności tych zmiennych.

Na wykładzie omówione zostało pięć miar współzmienności. Na początku dla każdego współczynnika napiszemy odpowiednią funkcję, która wylicza jego wartość na podstawie danych w postaci tabeli. Następnie dla badanych zmiennych zestawimy wszystkie współczynniki w tabeli i wyciągniemy odpowiednie wnioski, o ile testy niezależności pokażą, że hipotezę o niezależności można odrzucić.

(a) Współczynnik τ Goodmana.

```
tau <- function(t){
  R <- nrow(t)
  C <- ncol(t)
  n <- sum(t)
  K <- 0
  L <- 0
  for(i in 1:R){
    for(j in 1:C){
      K <- K+t[i,j]^2/(n*sum(t[i,]))
    }
  }
  for(j in 1:C){
    L <- L+(sum(t[,j])/n)^2
  }
  return((K-L)/(1-L))
}
```

(b) Współczynnik V Craméra.

```r
#na początku funkcja definiująca statystykę z testu chi2 Pearsona
X2 <- function(t){
  R <- nrow(t)
  C <- ncol(t)
  n <- sum(t)
  x2 <- 0
  for(i in 1:R){
    for(j in 1:C){
      x2 <- x2 + (t[i,j]-sum(t[i,])*sum(t[,j])/n)^2/(sum(t[i,])*sum(t[,j])/n)
    }
  }
  return(x2)
}

V <- function(t){
  R <- nrow(t)
  C <- ncol(t)
  n <- sum(t)
  v <- X2(t)/(n*min(R-1, C-1))
  return(sqrt(v))
}
```

(c) Współczynnik T Czuprowa.

```r
T.cz <- function(t){
  R <- nrow(t)
  C <- ncol(t)
  n <- sum(t)
  t.cz <- X2(t)/(n*sqrt((R-1)*(C-1)))
  return(sqrt(t.cz))
}
```

(d) Współczynnik φ.

```r
Phi <- function(t){
  n <- sum(t)
  phi <- X2(t)/n
  return(sqrt(phi))
}
```

(d) Współczynnik C Pearsona

```r
C.p <- function(t){
  n <- sum(t)
  c.p <- X2(t)/(X2(t)+n)
  return(sqrt(c.p))
}
```

3.1.1 Reakcja a miejsce

W tym przypadku mamy do czynienia z tablicą 2×2. Odpowiednie wartości prezentujemy w poniższej tabeli, gdzie wiersze odpowiadają zmiennej reakcja, a kolumny – zmiennej miejsce:

	dom	szpital
nie	86	61
tak	14	39

Tabela 8: Zestawienie zmiennych reakcja i miejsce

Do weryfikacji hipotezy o niezależności zmienych możemy użyć chi kwadrat Pearsona. Obserwujemy znikomo małą wartość poziomu krytycznego, więc hipotezę o niezależności odrzucamy z prawdopodobieństwem 1.

```
chisq.test(t1)$p.value

## [1] 0.0001204076
```

Odpowiednie miary współzmienności przedstawiamy w poniższej tabeli. Rzeczywiście, zgodnie z własnością tabeli 2×2, mamy $\tau = V^2 = T^2 = \varphi^2$. W interpretacji skupimy się więc na współczynniku τ, którego wartość jest bliska zeru, co świadczy o bardzo słabym powiązaniu badanych zmiennych, niemal rzeczywistej niezależności.

τ	V	T	φ	C
0.08022	0.28323	0.28323	0.28323	0.27251

Tabela 9: Miary współzmienności dla zmiennych reakcja, miejsce

3.1.2 Reakcja a dawka

Zbadamy teraz relację miedzy zmiennymi reakcja i dawka. Mamy do czynienia z tablicą 2×5 i podobnie jak poprzednio, odpowiednie wartości prezentujemy poniżej, gdzie wiersze odpowiadają zmiennej reakcja, a kolumny – zmiennej dawka (w skali logarytmicznej):

	-2	-2,301	-2,602	-2,903	-3,204
nie	21	25	32	32	37
tak	19	15	8	8	3

Tabela 10: Zestawienie zmiennych reakcja i dawka

Ponownie, uzywywamy testu chi kwadrat Pearsona do weryfikacji hipotezy niezależności i ponownie, otrzymujemy znikomo małe p-value, więc hipotezę o niezależności odrzucamy z prawdopodobieństwem 1.

```
chisq.test(t2)$p.value

## [1] 0.0003646877
```

Do obliczenia odpowiednich miar współzmiennności użyjemy funkcji zdefiniowanych poprzednio. Widzimy, że wartość współczynnika τ jest bliska zeru, jednak nie na tyle bliska, jak w przypadku zmiennych reakcja i miejsce. Na tej podstawie możemy stwierdzić, że – mimo wyników testu o niezależności – są one mniej zależne niż zmienne reakcja, dawka.

τ	V	T	φ	C
0.02586	0.32164	0.22743	0.32164	0.30619

Tabela 11: Miary współzmienności dla zmiennych reakcja, dawka

3.2 Zadanie 2

Na podstawie danych zawartych w poniższej tabelii obliczymy odpowiednią miarę współzmienności zmiennych `wynagrodzenie` oraz `stopień zadowolenia z pracy`.

	WYNAGRODZENIE	BN	N	Z	BZ
1	... - 6000	32	44	60	70
2	6000 - 15000	22	38	104	125
3	15000 - 25000	13	48	61	113
4	25000 - ...	3	18	54	96

Tabela 12: Wynagrodzenie badanych 901 osób i ich stopień zadowolenia

Użyjemy testu chi-kwadrat Pearsona w celu wyznaczenia $p - value$.

```
##
##  Pearson's Chi-squared test
##
## data:  z
## X-squared = 51.83, df = 9, p-value = 4.868e-08
```

Otrzymaliśmy bardzo małą wartość poziomu krytycznego, zatem odrzucamy hipotezę o niezależności zmiennych `wynagrodzenie` oraz `stopień zadowolenia z pracy`.

Do obliczenia miary współzmienności gamma skorzystamy z funkcji zdefiniowanej w zadaniu powyżej.

```
## [1] 0.2398433
```

Wykorzystaliśmy tutaj miarę gamma, ponieważ mamy do czynienia ze zmiennymi o uporządkowanych kategoriach.

Przeprowadzimy teraz analizę korespondencji, która jest opisową i eksploracyjną techniką dostarczająca nam informacji o strukturze powiązań między dwiema jakościowymi zmiennymi losowymi.

Żeby zbadać strukturę powiązań wynagrodzenia oraz stopnia zadowolenia z pracy przeanalizujemy empiryczne rozkłady warunkowe. Wyniki przedstawia poniższa tabela:

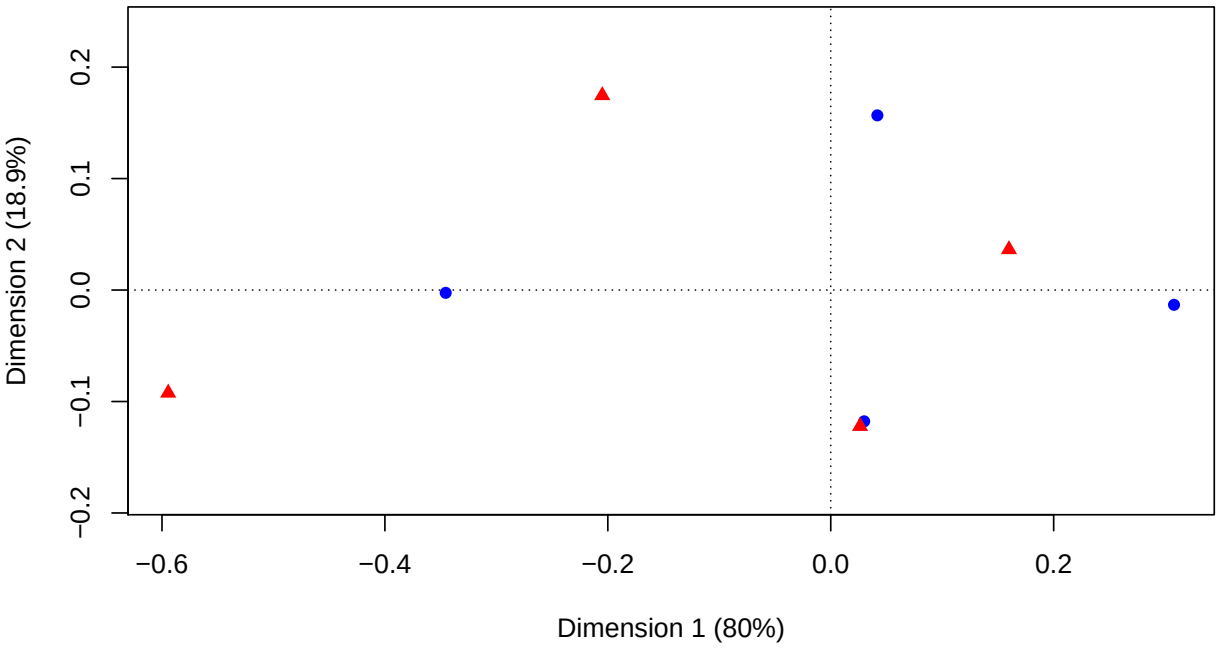
	WYNAGRODZENIE	BN	N	Z	BZ	SUMA
1	... - 6000	0,16	0,21	0,29	0,34	1
2	6000 - 15000	0,08	0,13	0,36	0,43	1
3	15000 - 25000	0,05	0,21	0,26	0,48	1
4	25000 - ...	0,02	0,10	0,32	0,56	1
5	Rozkład empiryczny	0,08	0,16	0,31	0,45	1

Tabela 13: Rozkłady empiryczne w wierszach

Z powyższej tabeli wynika, że odsetek osób bardzo niezadowolonych z pracy i zarabiających powyżej 25000 zł jest aż 8 razy niższy niż liczba osób bardzo niezadowolonych i zarabijących mniej niż 6000.

Graficzna prezentacja związku między stopniem zadowolenia z pracy a wysokością wynagrodzenia:

```
z <- matrix(c(32, 44, 60, 70, 22, 38, 104, 125, 13, 48, 61, 113, 3, 18, 54, 96),
            nrow = 4, byrow = TRUE)
plot(ca::ca(z))
```



Rysunek 1: Analiza korespondencji

3.3 Zadanie 3

200 klientów (w różnym wieku) kilku aptek zapytano, jaki lek przeciwbólowy zwykle stosują. Zebrane dane znajdują się w tabelce poniżej. Na podstawie tych danych obliczymy odpowiednie miary współzmienności oraz przeprowadzimy analizę korespondencji.

	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

Tabela 14: Dane dotyczące środków przeciwbólowych

Obliczymy miarę współzmienności τ Goodmana, ponieważ nie mamy tutaj sytuacji $2x2$:

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

Wyznaczymy teraz macierz korespondencji:

	do lat 35	od 36 do 55	powyżej 55
Ibuprom	0,35	0	0
Apap	0,22	0,22	0
Paracetamol	0,15	0,15	0,15
Ibuprofen	0	0,4	0,1
Panadol	0,18	0,03	0,05

Tabela 15: Macierz korespondencji

Oraz macierz profili wierszowych:

	do lat 35	od 36 do 55	powyżej 55
Ibuprom	1	0	0
Apap	0.5	0.5	0
Paracetamol	0,33	0,33	0,33
Ibuprofen	0	0,8	0,2
Panadol	0,7	0,12	0,18

Tabela 16: Macierz profili wierszowych

Widzimy, że aż 100% więcej osób w wieku do 35 lat używa Ibuprom zamiast Ibuprofenu. Wydaje się więc, że jest to jeden z najpopluarniejszych środków przeciwbólowych w tej grupie wiekowej. Zaraz po nim jest Panadol - również używa go duży odsetek osób.

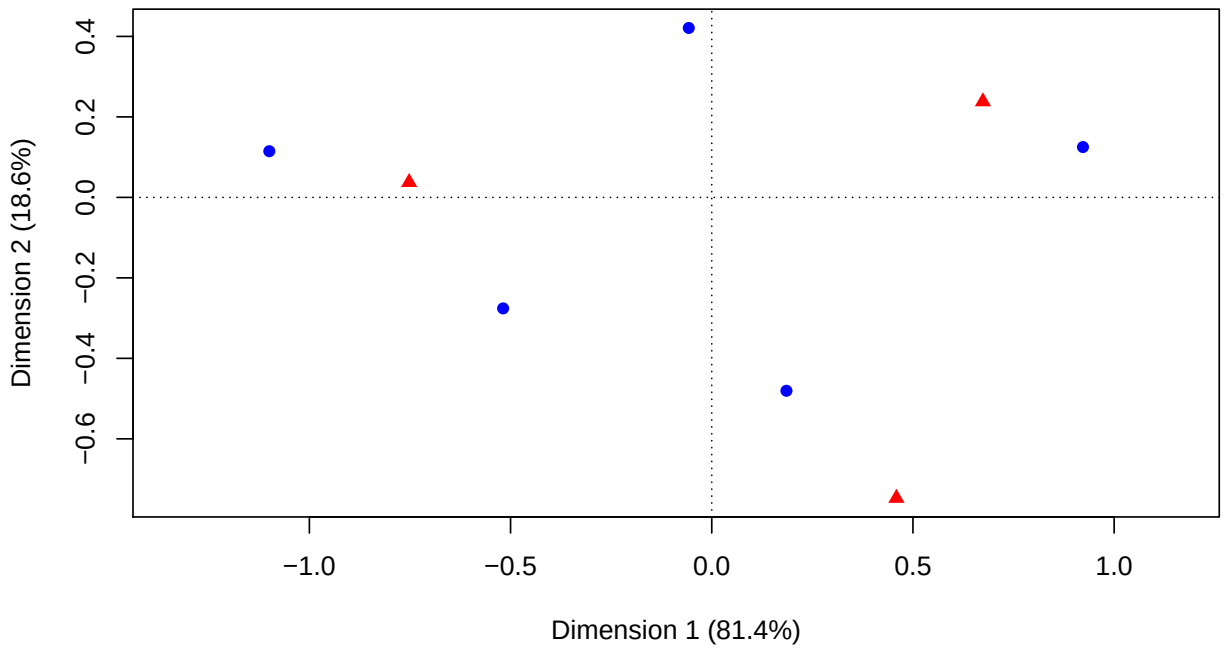
W następnym przedziale wiekowym sytuacja już się zmienia - w badanej grupie nikt nie używa już Ibupromu, króluje za to Ibuprofen.

W grupie osób powyżej 55 roku życia używane są już tylko: Paracetamol, Ibuprofen oraz Panadol.

Środkiem używanym przez wszystkie grupy wiekowe jest Paracetamol oraz Panadol.

Następnie korzystając z dostępnej w pakiecie `R` funkcji `ca` z biblioteki `ca` wyznaczymy odpowiednie wykresy oraz graficznie zaprezentujemy analizę korespondencji.

```
##
##  Principal inertias (eigenvalues):
##             1        2
## Value       0.467756 0.107084
## Percentage 81.37%   18.63%
##
##
##  Rows:
##               [,1]      [,2]      [,3]     [,4]      [,5]
## Mass     0.175000  0.220000  0.225000 0.250000  0.130000
## ChiDist  1.105542  0.424918  0.515201 0.930949  0.587300
## Inertia  0.213889  0.039722  0.059722 0.216667  0.044840
## Dim. 1  -1.607728 -0.083492  0.271406 1.348835 -0.758110
## Dim. 2   0.350706  1.286727 -1.468664 0.382309 -0.842934
##
##
##  Columns:
##               [,1]     [,2]      [,3]
## Mass     0.450000 0.400000  0.150000
## ChiDist  0.752962 0.714715  0.877058
## Inertia  0.255128 0.204327  0.115385
## Dim. 1  -1.099569 0.985364  0.671070
## Dim. 2   0.114764 0.727364 -2.283929
```



Rysunek 2: Analiza korespondencji