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 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.

Tablica 1: Dane do zadania 2. i 3.

	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

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
```

```r
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:

```r
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:

```r
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.

```r
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
```

```r
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.

```r
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.int = TRUE, 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.

```r
data2 <- matrix(c(18,72,3,77),nrow = 2, ncol = 2)
dimnames(data2) <- list(
```

```
Wiek_ankietowanych = c("Panadol", "Inny"),
Lek = c("do lat 35", "od 36 do 55"))
print(data2)

##                   Lek
## Wiek_ankietowanych do lat 35 od 36 do 55
##            Panadol        18           3
##            Inny           72          77

fisher.test(data2)$p.value

## [1] 0.001788538
```

Jak widzmy p-wartość jest mniejsza niż 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.

Tablica 2: Dane do zadania 4.

	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

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

## [1] 0.2139542
```

Jak możemy zobaczyć p-wartość jest większa niż 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:

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

Powyższa statystyka przy założeniu hipotezy o niezależności, dąży przy $n \to \infty$ do rozkładu χ^2 z $(R-1)(C-1)$ stopniami swobody, gdzie R i C są odpowiednią liczbą wierszy i kolumn testowanej tabeli.

```
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 niż 0.05, zatem nie mamy podstaw do odrzucenia hipotezy o niezależności rozkładów warunkowych.

Lista 8 i 9

Zad 1

Na podstawie danych z przykładu 1 na wykładzie 7, obliczyć wartości odpowiednich miar współzmienności zmiennych Segregacja (odpowiedź na pytanie dotyczace segregacji śmieci) i Wiek oraz Segregacja i Miejsce zamieszkania. W przypadku zmiennej Wiek, wartości miar obliczyć przy "wyjściowych" kategoriach wiekowych, jak również przy połączonych kategoriach wiekowych (jak w przykładzie). Podać interpretację tych wartości. Następnie przeprowadzić analizę korespondencji, tzn. obliczyć wartości odpowiednich macierzy, współrzędnych punktów oraz utworzyć odpowiednie wykresy.

Współczynnik τ Goodmana

$$\tau = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} p_{i,j}^2/p_{i+} - \sum_{j=1}^{C} p_{+j}^2}{1 - \sum_{j=1}^{C} p_{+j}^2}$$

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 $pj|i = 1$.

```
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) )
}
```

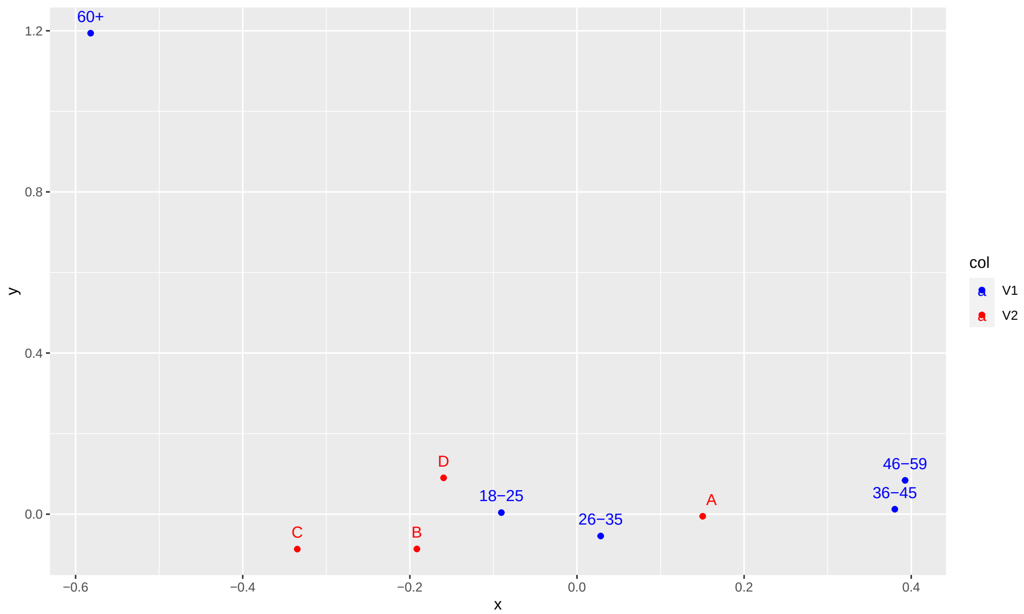
Tabela z przykładu 1 wykład 7:

```
##            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
## tau = 0.01712163
```

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

```
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)
}
df1 <-analiza(data4)
```

```
ggplot(df1, aes(x, y, label = name ,colour = col) ) +
        geom_point() + geom_text(vjust = -1) +
        scale_color_manual(values=c("#0000ff","#ff0000"))
```



Rysunek 1. Analiza korespondecji dla tabeli za Przykładu 1 Wykład 7.

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.

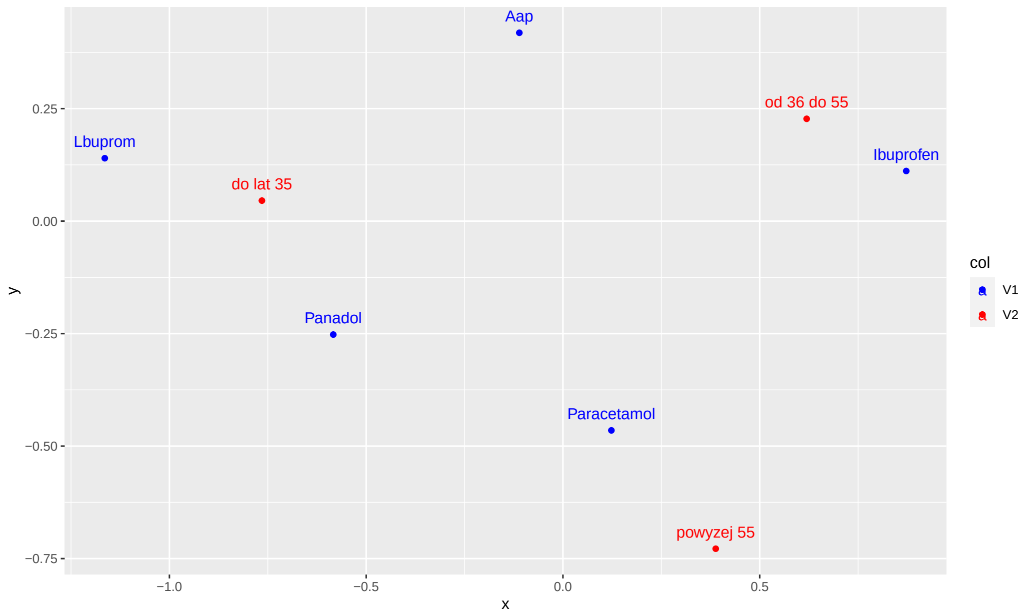
Tablica 3: Dane do zadania 2.

	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

```
## tau = 0.3194712
```

Jak możemy zobaczyć otrzymana wartość współczynnika τ nie jest zbyt duża, świadczy to o średniej współzmienności badanych zmiennych losowych.

```
df2 <-analiza(data5)
ggplot(df2, aes(x, y, label = name ,colour = col) ) +
        geom_point() + geom_text(vjust = -1) +
        scale_color_manual(values=c("#0000ff","#ff0000"))
```



Rysunek 2. Analiza korespondecji dla Tabeli 1.

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.

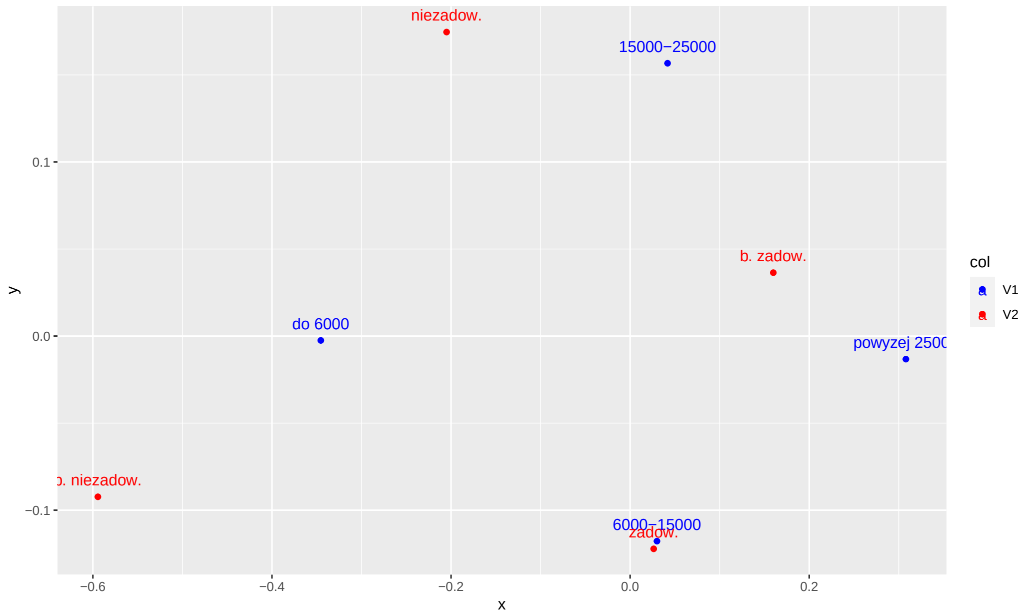
Tablica 4: Dane do zadania 3.

	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

```
## tau = 0.01671182
```

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

```
df3 <-analiza(data6)
ggplot(df3, aes(x, y, label = name ,colour = col) ) +
        geom_point() + geom_text(vjust = -1) +
        scale_color_manual(values=c("#0000ff","#ff0000"))
```



Rysunek 3. Analiza korespondecji dla Tabeli 2.