Analiza danych ankietowych

Sprawozdanie 3

Weronika Jaszkiewicz Weronika Pyrtak

Spis treści

Część I oraz II

Zadanie 1

Funkcja *p_wartosc_warunkowy_test_symetrii()* realizuje warunkowy test symetrii dla tabeli 2×2. Test opiera się na liczbie niesymetrycznych par, których suma traktowana jest jako próba w rozkładzie dwumianowym z prawdopodobieństwem sukcesu 0.5 (hipoteza symetrii). P-wartość obliczana jest jako dwustronne prawdopodobieństwo uzyskania wyniku co najmniej tak ekstremalnego jak zaobserwowany.

```
p_wartosc_warunkowy_test_symetrii<- function(tabela){
  n1 <- tabela[1,2]
  n2 <- tabela[2,1]
  n <- n1 + n2
  p <- 0
```

```r
  if(n1 < n/2){
    for (i in 1:n1) {
    p <- p + choose(n, i) * (0.5)^i * (0.5)^(n - i)
    }
  }else if(n1 > n/2){
    for (i in n1:n) {
    p <- p + choose(n, i) * (0.5)^i * (0.5)^(n - i)
    }
  }else{
      p <- 1
  }
    return(list(p_value = p))
  }
```

Zadanie 2

Dane dotyączcze reakcji na lek po godzinie od jego przyjęcia dla dwóch różnych leków przeciwbólowych stosowanych w migrenie zostały przedstawione w poniższej tabelce. Dla tych danych przeprowadzono test McNemara (z poprawką na ciągłość) oraz test warunkowy, miały one na celu zweryfikowanie hipotezy, że leki są jednakowo skuczene. Przyjmowany poziom istotności: $\alpha = 0.05$.

Tabela 1: Reakcja na lek A vs lek B

	Negatywna	Pozytywna
Negatywna	1	5
Pozytywna	2	4

Test McNemara z poprawka na ciagłość

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tabela_zad_2
## McNemar's chi-squared = 0.57143, df = 1, p-value = 0.4497
```

Wynik test wskazuje na brak podstaw do odrzucenia hipotezy zerowej. Oznacza to, że brak istotnych statystycznie różnic pomiędzy skutecznością leku A i leku B, zatem można uznać, że leki A i B są jednakowo skuteczne w tej próbie.

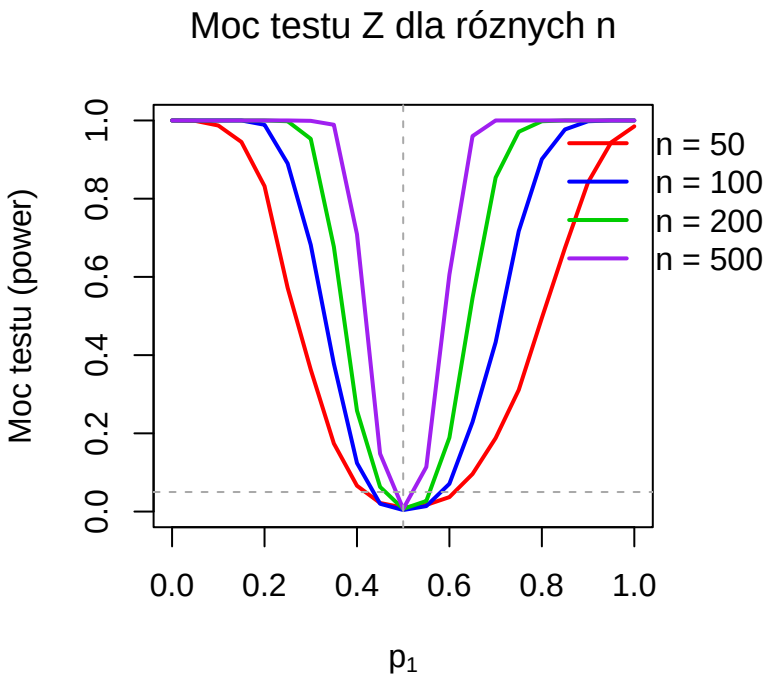
Test warunkowy

```
## $p_value
## [1] 0.2265625
```

P-wartość uzyskana w warunkowym teście symetrii jest znacznie większa od poziomu istotności. Oznacza to, że nie ma podstaw do odrzucenia hipotezy zerowej, czyli nie ma statystycznie istotnych różnic w skuteczności między lekiem A i lekiem B.

Zadanie 3

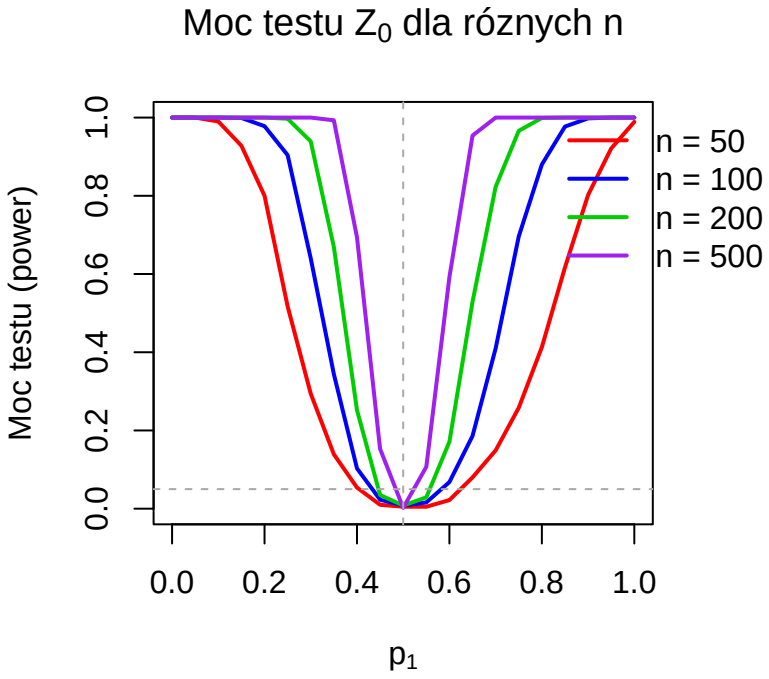
W celu porównania mocy testu Z oraz testu Z_0 przeprowadzono symulacje rozważając różne długości próby: $n = (50, 100, 200, 500)$.



Na wykresie przedstawiono estymowaną moc testu Z przy hipotezie zerowej $H_0 : p_1 = 0.5$. Krzywe mocy są symetryczne względem wartości $p_1 = 0.5$, co potwierdza, że test działa zgodnie z założeniem testowania dwustronnego.

Moc testu$Z$ wzrasta wraz z oddalaniem się wartości $p_1 = 0.5$. Oznacza to, że im większe jest rzeczywiste odchylenie od hipotezy zerowej, tym większa jest szansa na jej odrzucenie.

Z wykresu wynika również, że test Z staje się bardziej czuły wraz ze wzrostem liczności próby. Dla większych wartości moc testu szybciej rośnie i osiąga wartości bliskie 1. To wskazuje, że test jest bardziej skuteczny przy większych próbach.



Na wykresie przedstawiono estymowaną moc testu Z_0 przy hipotezie zerowej $H_0 : p_1 = 0.5$. Widać wyraźną symetrię względem $p_1 = 0.5$, co jets zgodne z założeniem testowania dwustronnego.

Można zauważyć, że moc testu rośnie wraz z oddalaniem się od od $p_1 = 0.5$. – im większa różnica między wartością rzeczywistą a wartością podaną w hipotezie zerowej, tym większa szansa na jej odrzucenie.

Dodatkowo, dla większych prób test Z_0 jest bardziej czuły – moc rośnie szybciej i szybciej zbliża się do wartości 1. Oznacza to, że test łatwiej wykrywa niewielkie różnice przy większej liczbie obserwacji.

Na podstawie symulacji stwierdzono, że testy Z i Z_0 wykazują bardzo podobne właściwości – moc obu testów rośnie wraz z licznością próby oraz oddalaniem się $p_1 = 0.5$. Oba testy są symetryczne względem $p_1 = 0.5$, co jest zgodne z założeniem testowania dwustronnego. Nie zaobserwowano istotnych różnic w mocy między testami, co sugeruje, że w analizowanych warunkach są równoważne pod względem skuteczności.

Zadanie 4

Celem badania było zweryfikowanie hipotezy, że zadowolenie ze szkoleń w pierwszym badanym okresie i w drugim badanym okresie pierwszego badania odpowiada modelowi symetrii.

Tabela 2: Tabela zadowolenia: pomiar 1 vs pomiar 2

	NIE	TAK
NIE	74	20
TAK	8	98

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tabela_czy_zadow
## McNemar's chi-squared = 4.3214, df = 1, p-value = 0.03764
```

Na podstawie wyniku testu McNemara (z poprawka na ciagłość) odrzucamy hipotezę zerową ($p - value = 0.03764 < \alpha = 0.05$). Zatem możemy stwierdzić, że poziom zadowolenia ze szkoleń uległ istotnej statystycznie zmianie między pierwszym a drugim okresem badania.

Zadanie 5

Na podstawie danych przedstawionych w poniższej tableli sprawdzono, czy odpowiedzi w pierwszym badanym okresie i w drugim okresie odpowiadają modelowi symetrii. W tym celu przeprowadzono dwa testy:

Tabela 3: Tabela reakcji

	-2	-1	0	1	2
-2	10	2	1	1	0
-1	0	15	1	1	0
0	1	1	32	6	0
1	0	0	1	96	3
2	1	1	0	1	26

Test Bowkera

```
##
##  McNemar's Chi-squared test
##
## data:  tabela
## McNemar's chi-squared = NaN, df = 10, p-value = NA
```

Wynik testu Bowkera daje spodziewany wynik p-value = NA. Jest on spowodowany tym, że w liczniku statystyki testowej obliczamy $n_{ij} + n_{ji}$, co powoduje dzielenie przez 0.

Tesr IW

```
## $statistic
## [1] 13.32669
##
## $p_value
## [1] 0.2059752
```

W teście IW p-wartość przekracza standardowy poziom istotności ($\alpha = 0.05$), co zonacza, że nie ma podstaw do odrzucenia hipotezy zerowej.Zatem test IW również nie wykazuje istotnych różnic między ocenami podejścia firmy w dwóch okresach.

W związku z tym, także na podstawie tego testu można stwierdzić, że ocena podejścia firmy do umożliwiania wdrażania wiedzy nie uległa istotnej zmianie.

Część III

Zadanie 6

W pewnym badaniu porównywano skuteczność dwóch metod leczenia: Leczenie A to nowa procedura, a Leczenie B to stara procedura. Przeanalizowano dane przedstawione w Tabeli 3 (wyniki dla całej grupy pacjentów) oraz w Tabelach 4 i 5 (wyniki w podgrupach ze względu na dodatkową zmienną) i odpowiedz na pytanie, czy dla danych występuje paradoks Simpsona.

```
# Dane
all <- matrix(c(117, 104, 177, 44), nrow = 2, byrow = TRUE,
              dimnames = list(c("Leczenie A", "Leczenie B"), c("Poprawa", "Brak")))

with_comorb <- matrix(c(17, 101, 2, 36), nrow = 2, byrow = TRUE,
                      dimnames = list(c("Leczenie A", "Leczenie B"), c("Poprawa", "Brak"

without_comorb <- matrix(c(100, 3, 175, 8), nrow = 2, byrow = TRUE,
                         dimnames = list(c("Leczenie A", "Leczenie B"), c("Poprawa", "Br

# Funkcja do obliczania skuteczności
effectiveness <- function(data) {
  round(data[, "Poprawa"] / rowSums(data), 3)
}

# Wyniki
eff_all <- effectiveness(all)
eff_with <- effectiveness(with_comorb)
eff_without <- effectiveness(without_comorb)

eff_all
```

```
## Leczenie A Leczenie B
##      0.529      0.801
```

```
eff_with
```

```
## Leczenie A Leczenie B
##      0.144      0.053
```

```
eff_without
```

```
## Leczenie A Leczenie B
##      0.971      0.956
```

```r
# Testy chi-kwadrat niezależności
test_all <- chisq.test(all)
test_with <- chisq.test(with_comorb)
test_without <- chisq.test(without_comorb)

# Zbiór wyników
test_results <- data.frame(
  Tabela = c("Cała grupa", "Z chorobami", "Bez chorób"),
  Chi2 = round(c(test_all$statistic, test_with$statistic, test_without$statistic), 2),
  DF = c(test_all$parameter, test_with$parameter, test_without$parameter),
  p_value = round(c(test_all$p.value, test_with$p.value, test_without$p.value), 4)
)

test_results
```

```
##        Tabela  Chi2 DF p_value
## 1  Cała grupa 35.36  1  0.0000
## 2 Z chorobami  1.47  1  0.2248
## 3  Bez chorób  0.09  1  0.7675
```

Analiza skuteczności metod leczenia

Dla całej grupy pacjentów skuteczność leczenia wynosi:

$$\text{Leczenie A:} \quad \frac{117}{117+104} \approx 0{,}529$$
$$\text{Leczenie B:} \quad \frac{177}{177+44} \approx 0{,}801$$

Dla pacjentów z chorobami współistniejącymi:

$$\text{Leczenie A:} \quad \frac{17}{17+101} \approx 0{,}144$$
$$\text{Leczenie B:} \quad \frac{2}{2+36} \approx 0{,}053$$

Dla pacjentów bez chorób współistniejących:

$$\text{Leczenie A:} \quad \frac{100}{100+3} \approx 0{,}971$$
$$\text{Leczenie B:} \quad \frac{175}{175+8} \approx 0{,}956$$

Wniosek

Tabela	Statystyka χ^2	DF	p-value
Cała grupa	47.06	1	<0.0001
Z chorobami	1.19	1	0.2755
Bez chorób	0.18	1	0.6699

Tabela 4: Wyniki testów χ^2 niezależności dla skuteczności leczenia

W każdej podgrupie (zarówno pacjentów z chorobami współistniejącymi, jak i bez nich) leczenie A okazuje się skuteczniejsze niż leczenie B. Jednakże w całej populacji obserwujemy odwrotny wniosek — leczenie B ma wyższą skuteczność. Jest to klasyczny przykład paradoksu Simpsona, w którym agregacja danych zaciemnia rzeczywiste zależności występujące w podgrupach.

Dla całej grupy różnica skuteczności między Leczeniem A i B jest statystycznie istotna ($p < 0.0001$).

W podgrupach (z i bez chorób współistniejących) nie ma podstaw do odrzucenia hipotezy niezależności – brak istotnych różnic w skuteczności między metodami. To potwierdza występowanie paradoksu Simpsona – agregacja danych prowadzi do innych wniosków niż analiza w podgrupach.

Zadanie 7

Dla danych z listy 1, przyjmując za zmienną 1 zmienną CZY_KIER, za zmienn 2– zmienną PYT_2 i za zmienną 3– zmienną STAŻ, przedstawiono interpretacje nastepujacych modeli log-liniowych: [13], [13], [123], [123], [1213] oraz [123].

```
# Zakładamy, że dane masz w ramce danych `dane`
# zmienne: CZY_KIER, PYT_2, STAZ
# Wczytanie danych
dane <- read.csv("ankieta.csv", sep = ";", fileEncoding = "Latin2")
colnames(dane) <- c('DZIAŁ','STAZ','CZY_KIER', 'PYT_1', 'PYT_2', 'PYT_3', 'PLEC', 'WIEK'

# Tabela kontyngencji 3D
tablica <- xtabs(~ CZY_KIER + PYT_2 + STAZ, data = dane)

# powinno zwrócić: "CZY_KIER" "PYT_2" "STAZ"
library(MASS)

# Lista nazw i formuł modeli log-liniowych
model_names <- c("[1 3]", "[13]", "[1 2 3]", "[12 3]", "[12 13]", "[1 23]")
formulas <- list(
  ~ CZY_KIER + STAZ,
```

```
  ~ CZY_KIER + STAZ + CZY_KIER:STAZ,
  ~ CZY_KIER * PYT_2 * STAZ,
  ~ CZY_KIER * PYT_2 + STAZ,
  ~ CZY_KIER * PYT_2 + CZY_KIER * STAZ,
  ~ CZY_KIER + PYT_2 * STAZ + CZY_KIER:STAZ
)

# Dopasowanie modeli i zapis wyników
results <- data.frame(Model = model_names, Deviance = NA, DF = NA, p_value = NA)

for (i in seq_along(formulas)) {
  fit <- loglm(formulas[[i]], data = tablica)
  results$Deviance[i] <- round(fit$deviance, 2)
  results$DF[i] <- fit$df
  results$p_value[i] <- round(pchisq(fit$deviance, df = fit$df, lower.tail = FALSE), 4)
}

results
```

Model	Deviance	DF	p-value
[1 3]	203.07	20	0.0000
[13]	183.98	18	0.0000
[1 2 3]	0.00	0	1.0000
[12 3]	33.91	14	0.0021
[12 13]	14.82	12	0.2512
[1 23]	4.88	9	0.8446

Tabela 5: Dopasowanie modeli log-liniowych: wartość statystyki deviance, liczba stopni swobody i wartość p.

Na podstawie analizy modeli log-liniowych można stwierdzić, że najlepszym dopasowaniem do danych charakteryzuje się model [1 23], który uwzględnia zależność pomiędzy zmiennymi PYT_2 i STAZ oraz ich wspólny wpływ na CZY_KIER. Model ten ma wysoką wartość p-value (0,8446), co oznacza brak podstaw do jego odrzucenia, a jednocześnie jest prostszy niż model pełny [1 2 3]. Modele [1 3] i [13] należy odrzucić ze względu na istotnie słabe dopasowanie ($p < 0{,}001$).

Zadanie 8

Przyjmując model log-liniowy [123] dla zmiennych opisanych w zadaniu 7 oszacowano prawdopobiebieństwa:

- ze osoba pracuj ¸ aca na stanowisku kierowniczym jest zdecydowanie zadowolona ze szkoleń,

- ze osoba o staż pracy krótszym niż rok pracuje na stanowisku kierowniczym;
- ze osoba o stażu pracy powyżej trzech lat nie pracuje na stanowisku kierowniczym.

Jakie byłyby oszacowania powyższych prawdopodobieństw przy założeniu modelu [1223]?

Sytuacja	Prawdopodobieństwo
1. Osoba na stanowisku kierowniczym, zdecydowanie zadowolona ze szkoleń (PYT_2 = "2")	0.1667
2. Osoba o stażu krótszym niż 1 rok (STAŻ = "1"), nie pracuje na stanowisku kierowniczym	0.2083
3. Osoba o stażu powyżej 3 lat (STAŻ = "3"), nie pracuje na stanowisku kierowniczym	0.0833

Tabela 6: Prawdopodobieństwa przy założeniu modelu log-liniowego [1 2 3]

Sytuacja	Prawdopodobieństwo
1. Osoba na stanowisku kierowniczym, zdecydowanie zadowolona ze szkoleń (PYT_2 = "2")	0.1513
2. Osoba o stażu krótszym niż 1 rok (STAŻ = "1"), nie pracuje na stanowisku kierowniczym	0.2174
3. Osoba o stażu powyżej 3 lat (STAŻ = "3"), nie pracuje na stanowisku kierowniczym	0.0865

Tabela 7: Prawdopodobieństwa przy założeniu modelu log-liniowego [12 23]

Prawdopodobieństwa oszacowane przez oba modele są do siebie bardzo zbliżone. Model pełny [123] odwzorowuje dokładnie strukturę danych - jest nadmiernie dopasowany, natomiast model [1223] daje podobne wyniki przy mniejszej liczbie interakcji, dlatego może być uznany za bardziej parsymonialny i praktyczny w interpretacji.

Część II

Zadanie 9

Dla danych wskazanych w zadaniu 7 zweryfikowano następujące hipotezy:

- zmienne losowe CZY_KIER, PYT_2 i STAŻ sa wzajemnie niezależne;
- zmienna losowa PYT_2 jest niezależna od pary zmiennych CZY_KIER i STAŻ;
- zmienna losowa PYT_2 jest niezależna od zmiennej CZY_KIER, przy ustalonej wartości zmiennej STAŻ

```
##                                   Hipoteza  Deviance DF p_value
## 1               H1: całkowita niezależność 42.242215 17  0.0006
## 2 H2: PYT_2 niezależna od (CZY_KIER, STAZ) 23.152114 15  0.0810
## 3              H3: PYT_2  CZY_KIER | STAZ  4.879959  9  0.8446
```

Hipoteza	Deviance	DF	p-value
H1: całkowita niezależność (CZY_KIER, PYT_2, STAZ)	42.24	17	0.0006
H2: PYT_2 niezależna od pary (CZY_KIER, STAZ)	23.15	15	0.0810
H3: PYT_2 $\perp$ CZY_KIER | STAZ (warunkowa niezależność)	4.88	9	0.8446

Tabela 8: Weryfikacja hipotez o niezależności między zmiennymi za pomocą modeli log--liniowych

Hipoteza H1 (całkowita niezależność wszystkich trzech zmiennych) została odrzucona na poziomie istotności 0,05 — bardzo niskie p-value (0.0006) świadczy o silnych zależnościach między zmiennymi.

Hipoteza H2 (PYT_2 niezależna od pary CZY_KIER i STAZ) nie została odrzucona, ale wartość p = 0.0810 jest bliska granicy — wskazuje na możliwy umiarkowany związek.

Hipoteza H3 (warunkowa niezależność PYT_2 CZY_KIER przy ustalonym STAZ) nie została odrzucona — wysokie p-value (0.8446) sugeruje, że warunkowa niezależność jest uzasadniona i dobrze opisuje dane.