3

Lista 3

Część I i II

Zadanie 1

Napisz funkcję, która zwraca p-wartość w omówionym na wykładzie warunkowym teście symetrii w przypadku tabeli 2×2.

```
p <- function(n12, n21){
  part <- 0
  if(n12<(n12+n21)/2){
    for(i in 0:n12){
    part <- part + choose(n12+n21,i)*(1/2)^i*(1/2)^(n12+n21-i)
    }
    part <- 2*part
  }
  if(n12>(n12+n21)/2){
    for(i in 0:n12+n21){
    part <- part + choose(n12+n21,i)*(1/2)^i*(1/2)^(n12+n21-i)
    }
    part <- 2*part
  }
  if(n12==(n12+n21)/2){
    part <- 1
  }
  return(part)
}
```

Zadanie 2

Zadanie 2.1

```r
tabela <- matrix(c(1, 2, 5, 4), nrow = 2,
dimnames = list("Lek A" = c("Negatywna", "Pozytywna"),
"Lek B" = c("Negatywna", "Pozytywna")))
print(tabela)
```

```
           Lek B
Lek A       Negatywna Pozytywna
  Negatywna         1         5
  Pozytywna         2         4
```

```r
#test McNemara
mcnemar.test(tabela, correct = TRUE)
```

```

	McNemar's Chi-squared test with continuity correction

data:  tabela
McNemar's chi-squared = 0.57143, df = 1, p-value = 0.4497
```

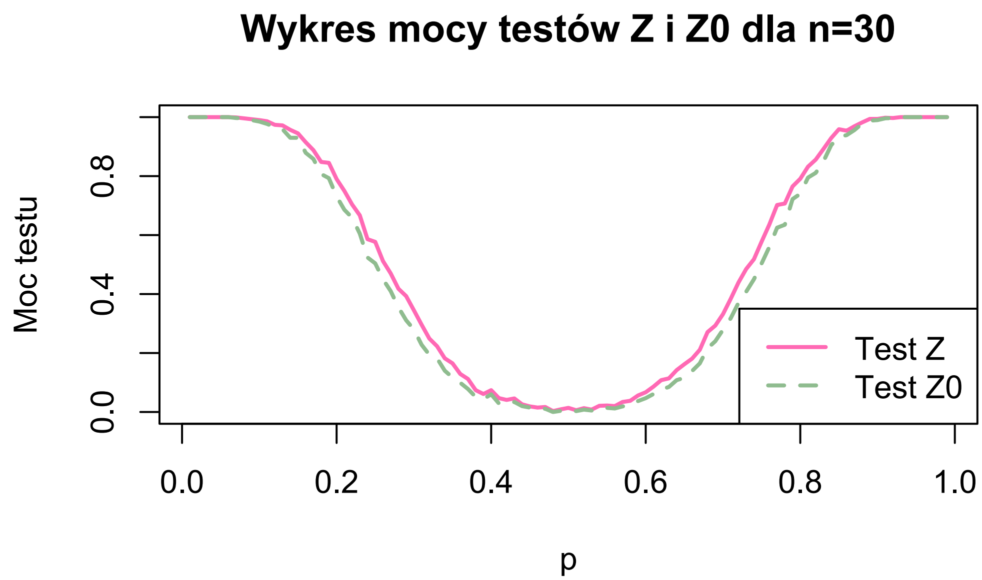
Zadanie 2.2

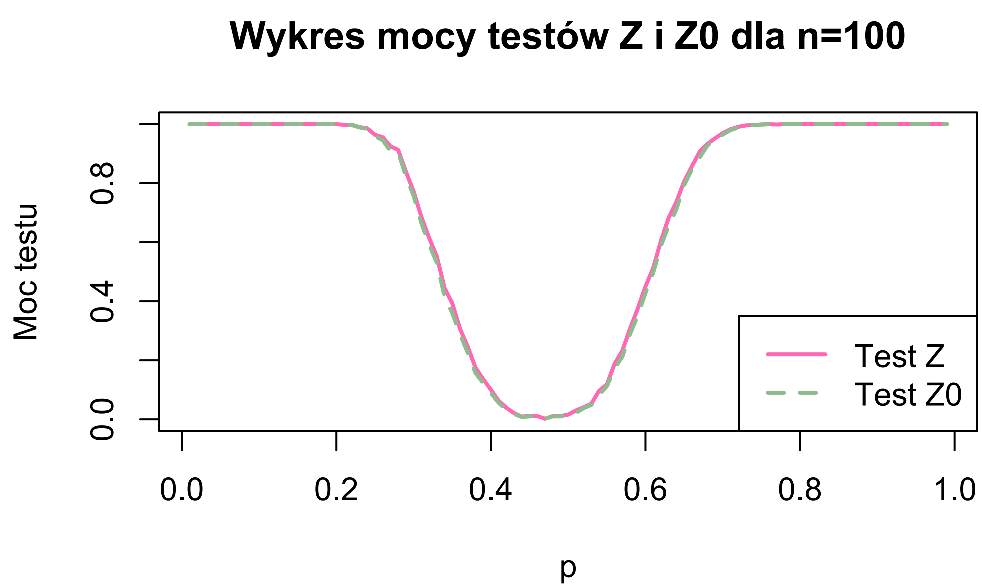
```
[1] 0.453125
```

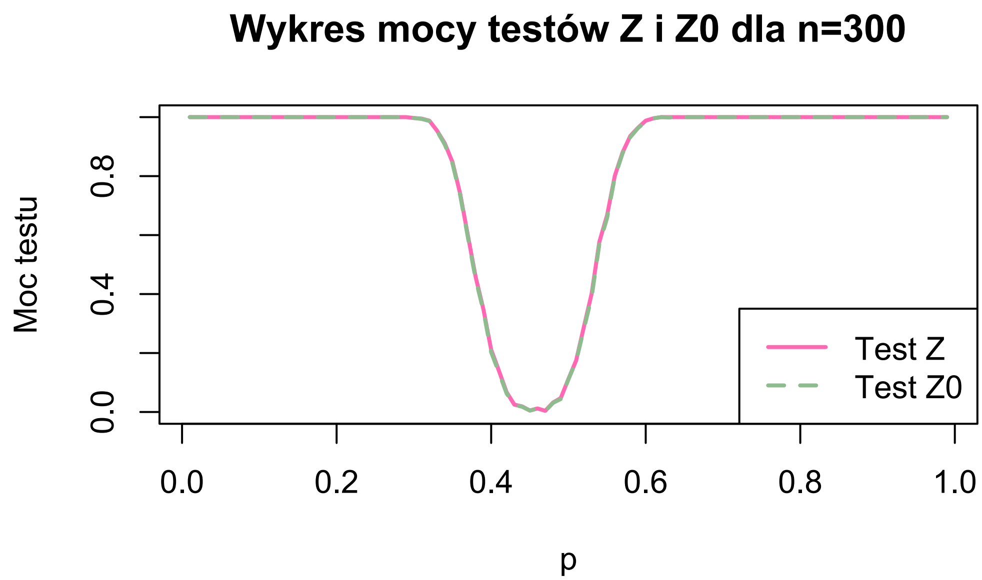
```
[1] 0.453125
```

Zadanie 3

Przeprowadź symulacje w celu porównania mocy testu Z i testu Z_0 przedstawionych na wykładzie. Rozważ różne długości prób.


Wykres mocy testów Z i Z0 dla n=30
Moc testu
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
p
Test Z
Test Z0


Wykres mocy testów Z i Z0 dla n=100
Moc testu
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
p
Test Z
Test Z0



Widzimy, że dla mniejszych n test Z ma większą moc od testu Z_0. Dla większych n moce testów zbliżają sie do siebie oraz rosną, szczególnie wokół $p = 0.5$.

Zadanie 4

```
    McNemar's Chi-squared test with continuity correction

data:  tabela
McNemar's chi-squared = 4.3214, df = 1, p-value = 0.03764
```

Część III

Zadanie 6

W pewnym badaniu porównywano skuteczność dwóch metod leczenia: Leczenie A to nowa procedura, a Leczenie B to stara procedura. Przeanalizuj 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.

Table 1: Tabela 3: Dane dla całej grupy

Metoda	Poprawa	Brak poprawy
Leczenie A	117	104
Leczenie B	177	44

Table 2: Tabela 4: Dane dla pacjentów z chorobami współistniejącymi.

Metoda	Poprawa	Brak poprawy
Leczenie A	17	101
Leczenie B	2	36

Table 3: Tabela 5: Dane dla pacjentów bez chorób współistniejących.

Metoda	Poprawa	Brak poprawy
Leczenie A	100	3
Leczenie B	175	8

```
    wszyscy z chorobamu bez chorób
A 0.5294118  0.14406780  0.9708738
B 0.8009050  0.05263158  0.9562842
```

Chociaż leczenie B "wygrywa" patrząc na całą grupę badanych, po podziale na grupy ze względu na obecność chorób współistniejących możemy zauważyć, że to leczenie A ma większy odsetek wyzdrowień.

```
     wszyscy  z chorobamu   bez chorób
2.740007e-09 2.248419e-01 7.675118e-01
```

W przeprowadzonym teście niezależności χ^2 dla całej grupy p-value jest bardzo małe, więc odrzucamy hipotezę H_0 o niezależności. Jednak ten sam test wykonany osobno dla badanych grup - z chorobami współistniejącymi oraz bez chorób - w obu przypadkach daje p-value większą od poziomu istotności, a więc nie mamy podstaw do odrzucania hipotezy zerowej o niezależności zmiennych, to znaczy wyniku leczenia (poprawy) od przyjętego leczenia. To znaczy, że pozorny związek dla całej badanej grupy nie przekłada się na zaleznośź w podgrupach - a więc jest to klasyczny przypadek paradoksu Simpsona.

Zadanie 7

Dla danych z listy 1, przyjmując za zmienną 1 zmiennaą CZY_KIER, za zmienną 2 – zmienną PYT_2 i za zmienną 3 – zmienną STAŻ, podaj interpretacje następujących modeli log-liniowych: [1 3], [13], [1 2 3], [12 3], [12 13] oraz [1 23].

[1 3] - zmienne CZY_KIER oraz STAŻ są niezależne,

[13] - zmienne CZY_KIER oraz STAŻ nie są niezależne,

[1 2 3] - zmienne CZY_KIER, PYT_2 oraz STAŻ są niezależne,

[12 3] - zmienne CZY_KIER i PYT_2 nie są niezależne, a zmienna STAŻ jest niezależna od nich obu,

[12 13] - zmienne CZY_KIER i PYT_2 nie są niezależne, CZY_KIER i STAŻ nie są niezalezne, a PYT_2 i STAŻ są warunkowo niezależne,

[1 23] - zmienna CZY_KIER jest niezależna od pozostałych dwóch, PYT_2 i STAŻ, które nie są od siebie niezależne.

Część IV i V

Zadanie 8

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

- że osoba pracująca na stanowisku kierowniczym jest zdecydowanie zadowolona ze szkoleń;
- że osoba o stażu pracy krótszym niż rok pracuje na stanowisku kierowniczym;
- że osoba o stażu pracy powyżej trzech lat nie pracuje na stanowisku kierowniczym.

Jakie byłyby oszacowania powyższych prawdopodobieństw przy założeniu modelu [12 23]?

Zaczynamy od modelu [123]:

```
# A tibble: 4 x 5
  PYT_2 freq_sum fitted_sum p_dane p_model
  <fct>    <int>      <dbl>  <dbl>   <dbl>
1 -2          10      10.0  0.370   0.370
2 -1           2       2.00 0.0741  0.0741
3 1            2       2.00 0.0741  0.0741
4 2           13      13.0  0.481   0.481
```

```
# A tibble: 2 x 5
  CZY_KIER freq_sum fitted_sum p_dane p_model
  <fct>       <int>      <dbl>  <dbl>   <dbl>
1 Nie            40      40.0  0.976   0.976
2 Tak             1       1.00 0.0244  0.0244
```

```
# A tibble: 2 x 5
  CZY_KIER freq_sum fitted_sum p_dane p_model
  <fct>       <int>      <dbl>  <dbl>   <dbl>
1 Nie            10      10.0   0.526   0.526
2 Tak             9       9.00  0.474   0.474
```

Model [123] dobrze oszacował potrzebne prawdopodobieństwa (w 1. tabeli interesuje nas wiersz z odpowiedzią "2" na PYT_2, w 2. i 3. odpowiedź "Tak" w kolumnie CZY_KIER). Zarówno szacowane liczności jak i prawdopodobieństwa są równe dla modelu i danych.

```
# A tibble: 4 x 5
  PYT_2 freq_sum fitted_sum p_dane p_model
  <fct>    <int>      <dbl>  <dbl>   <dbl>
1 -2          10      10.0  0.370   0.370
2 -1           2       2.00 0.0741  0.0741
3 1            2       2.00 0.0741  0.0741
4 2           13      13.0  0.481   0.481
```

```
# A tibble: 2 x 5
  CZY_KIER freq_sum fitted_sum p_dane p_model
  <fct>       <int>      <dbl>  <dbl>   <dbl>
1 Nie            40      35.7  0.976    0.872
2 Tak             1       5.25 0.0244   0.128
```

```
# A tibble: 2 x 5
  CZY_KIER freq_sum fitted_sum p_dane p_model
  <fct>       <int>      <dbl>  <dbl>   <dbl>
1 Nie            10      14.8   0.526   0.778
2 Tak             9       4.22  0.474   0.222
```

Dla modelu [12 23] odpowiedź na pierwszy podpunkt się zgadza - wartości w danych są równe przewidywanym przez model. Jednak przy pytaniach, które łączą zmienne CZY_KIER oraz STAŻ (podpunkt 2. i 3.) model przeszacował wyniki z dla osób o krótkim stażu oraz niedoszacował odpowiedzi w dla osób o długim stażu - wynika to z braku powiązanie między tymi zmiennymi. Jak widzimy, złe dobranie modelu skutkuje złym oszacowaniem badanych prawdopodobieństw.

Zadania dodatkowe

Zadanie 2*

Na podstawie danych z listy 1 dokonaj wyboru modelu rozważając uwzględnienie zmiennych PYT_1, PYT_2 i PŁEĆ w oparciu o:

- testy,
- kryterium AIC,
- kryterium BIC.

Będzimy roważać modele [**1 2 3**], [**12 13 23**] oraz [**123**].

```
Analysis of Deviance Table

Model 1: Freq ~ PYT_1 + PYT_2 + PŁEĆ
Model 2: Freq ~ PYT_1 * PYT_2 + PYT_1 * PŁEĆ + PYT_2 * PŁEĆ
Model 3: Freq ~ PYT_1 * PYT_2 * PŁEĆ
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        31    226.057
2        12      7.365 19  218.692   <2e-16 ***
3         0      0.000 12    7.365   0.8326
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

W tej analizie testujemy, czy model prostszy wystarcza (H_0), czy potrzebny jest model bardziej złożony (H_1). Testujemy, czy sensownie jest zmieniać model z [1 2 3] na [12 13 23] oraz [12 13 23] na [123]. Wykonujemy test istotności χ^2.

Interpretacja:

Model 1 (niezależność) jest zbyt prosty, ponieważ po dodaniu interakcji dwójkowych (Model 2) dopasowanie znacznie się poprawia (p $<$ 2e-16 więc odrzucamy H_0).

Model 3 (pełny) nie poprawia istotnie dopasowania względem Modelu 2 (p = 0.8326, nie ma podstaw do odrzucenia H_0), więc interakcja trójkowa nie jest potrzebna.

Ostateczny wybór: Model 2 – zawiera wszystkie istotne interakcje (dwójkowe), a jest prostszym modelem niż pełny.

```
           Model      AIC      BIC
1     model_full 150.1856 217.7408
2 model_12_13_23 133.5509 180.8396
3    model_indep 314.2426 329.4425
```

Porównując **AIC** orac **BIC** widzimy, że dla obu kryteriów model [12 13 23] przymuje najmniejsze wartości, więc dla tego porównania jest najlepszy.

Zarówno testy chi-kwadrat, jak i kryteria AIC/BIC wskazują, że najlepszym modelem jest model z interakcjami dwójkowymi: `Freq ~ PYT_1*PYT_2 + PYT_1*PŁEĆ + PYT_2*PŁEĆ`, oznaczany jako [**12 13 23**].