Spis treści

1 Wstęp

Celem niniejszego sprawozdania jest przeprowadzenie analizy danych ankietowych dotyczących oceny szkoleń przeprowadzonych w firmie, a także badania zależności pomiędzy różnymi zmiennymi demograficznymi a opiniami pracowników. W ramach prac wykonamy szereg zadań obejmujących:

- wyznaczanie przedziałów ufności dla prawdopodobieństw opisujących poziom zadowolenia ze szkolenia,

- konstruowanie funkcji do wyznaczania poziomów krytycznych dla różnych testów statystycznych,

- weryfikację hipotez dotyczących rozkładów odpowiedzi oraz niezależności zmiennych za pomocą testów chi-kwadrat, Fishera oraz Freemana-Haltona,

- analizę wyników testów i ich graficzną interpretację przy użyciu wykresów asocjacyjnych,

- przeprowadzenie symulacji w celu oceny mocy testów statystycznych,

- ocenę zależności pomiędzy zmiennymi przy użyciu miar takich jak ryzyko względne, iloraz szans, współczynniki korelacji dla zmiennych porządkowych oraz analiza korespondencji.

Dla przejrzystości i uporządkowania analiz, raport został podzielony na pięć części, z których każda odpowiada kolejnym zagadnieniom badawczym. Sprawozdanie ma na celu rozwinięcie praktycznych umiejętności w zakresie stosowania metod statystycznych w analizie danych ankietowych oraz interpretacji uzyskanych wyników w kontekście problemów rzeczywistych.

2 Część I

W pierwszej części sprawozdania skupimy się na analizie danych dotyczących opinii pracowników na temat skuteczności szkolenia “Efektywna komunikacja w zespole”. Na podstawie odpowiedzi wyznaczymy przedziały ufności dla wektora prawdopodobieństw opisującego stopień zadowolenia ze szkolenia. Następnie przygotujemy funkcje umożliwiające wyznaczanie poziomów krytycznych w testach chi-kwadrat Pearsona i największej wiarygodności, a także wykorzystamy je do weryfikacji hipotezy o równomierności rozkładu odpowiedzi na pytanie dotyczące wsparcia i materiałów szkoleniowych w Dziale Produktowym. W analizie przyjmiemy poziomy istotności wskazane w treści zadań.

2.1 Zadanie 1

W ankiecie przedstawionej na poprzedniej liście pracownicy zostali poproszeni o wyrażenie opinii na temat skuteczności szkolenia "Efektywna komunikacja w zespole" zorganizowanego przez firmę. Wśród próbki 200 pracowników (losowanie proste ze zwracaniem) uzyskano wyniki:

- 14 pracowników- bardzo niezadowolonych,
- 17 pracowników- niezadowolonych,
- 40 pracowników- nie ma zdania,
- 100 pracowników- zadowolonych,
- 29 pracowników- bardzo zadowolonych,

Na podstawie danych wyznacz przedział ufności dla wektora prawodobieństw opisującego stopieN zadowolenia ze szkolenia. Przyjmij poziom ufności **0.95**.

Rozwiązanie

```
$Clopper_Pearson
           [,1]      [,2]
[1,] 0.03169652 0.1298937
[2,] 0.04208141 0.1486579
[3,] 0.13257329 0.2821753
[4,] 0.40735190 0.5926481
[5,] 0.08749866 0.2200467

$Wilson
           [,1]      [,2]
[1,] 0.03604773 0.1315662
[2,] 0.04660626 0.1500444
[3,] 0.13731215 0.2819534
[4,] 0.41040470 0.5895953
[5,] 0.09228421 0.2205134

$Wald
           [,1]      [,2]
[1,] 0.02352787 0.1164721
[2,] 0.03420487 0.1357951
[3,] 0.12714455 0.2728555
[4,] 0.40893068 0.5910693
[5,] 0.08086883 0.2091312
```

W zadaniu wyznaczyłyśmy przedziały ufności dla prawdopodobieństw opisujących stopień zadowolenia pracowników z przeprowadzonego szkolenia. Aby to osiągnąć:

- wykorzystałyśmy funkcję ***binom.confint()***,
- obliczenia przeprowadziłyśmy dla trzech różnych metod: Clopper-Pearson (dokładna metoda), Wilson oraz Wald (asymptotyczna metoda),
- dla każdej kategorii odpowiedzi obliczyłyśnmy osobno dolną i górną granicę przedziału ufności,
- wyniki przedstawione są w formie tabelarycznej oddzielnie dla każdej z metod.

Opis wyników

- Metoda Clopper-Pearson daje nam najszersze przedziały ufności, co wynika z jej charakteru — zapewnia większe bezpieczeństwo przy niskiej liczbie sukcesów lub porażek,
- Metoda Wilsona daje lekko węższe przedziały niż Clopper-Pearson, ale nadal zachowuje dobrą dokładność,
- Metoda Walda generuje najwęższe przedziały, ale ich dokładność dla małych lub skrajnych wartości może być niska.

Wnioski

Wyniki różnią się w zależności od wybranej metody. Metoda Clopper-Pearson jest najbardziej ostrożna (dłuższe przedziały), metoda Wilsona pozwala uzyskać przedziały węższe, przy zachowaniu wysokiej dokładności, natomiast metoda Walda daje najwęższe przedziały, ale ich wiarygodność może być ograniczona, zwłaszcza przy małych licznościach. W praktyce, dla wysokiej pewności wyników, zaleca się stosowanie metody Clopper-Pearson lub Wilsona.

2.2 Zadanie 2

Napisz funkcję, która wyznacza wartość poziomu krytycznego w następujących testach:

- chi-kwadrat Pearsona,
- chi-kwadrat największej wiarogodności,

służących do weryfikacji hipotezy H_0

```
test <- function(x, n, p0, alpha = 0.05) {
  statystyka1 <- sum((x - n * p0)^2 / (n * p0))
  p_val <- 1 - pchisq(statystyka1, length(p0)-1)

  statystyka2 <- 2 * sum(x * log(x / (n * p0)))
  p_val2 <- 1 - pchisq(statystyka2, length(p0)-1)
```

```r
  return(c(p_val, p_val2))
}


x <- c(20, 30, 40, 50)
n <- sum(x)
p0 <- c(0.2, 0.2, 0.2, 0.2)


test(x, n, p0, alpha)
```

```
[1] 1.653969e-05 1.110223e-16
```

2.3 Zadanie 3

Na podstawie danych z ankiety z poprzedniej listy zweryfikuj hipotezę, że w grupie pracowników zatrudnionwych w Dziale Produktowym rozkład odpowiedzi na pytanie "Jak bardzo zgadzasz się ze stwierdzeniem, że firma zapewnia odpowiednie wsparcie i materiały umożliwiające skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń?" jest równomierny, tzn. jest jednakowe prawdopodobieństwo, że pracownik zatrudniony w Dziale Produkcjnym udzielił odpowiedzi "zdecydowanie się nie zgadzam", "nie zgadzam się", "nie mam zdania", "zgadzamsię", "zdecydowanie się zgadzam" na pytanie PYT_1. Przyjmij poziom istotności 0.05. Skorzystaj z funkcji napisanej w zadaniu 2

3 Część II

W drugiej części raportu zajmiemy się badaniem zależności pomiędzy wybranymi zmiennymi ankietowymi. W szczególności zweryfikujemy hipotezy o niezależności zmiennych takich jak płeć, wiek, staż pracy i zajmowane stanowisko. W analizach wykorzystamy test Fishera oraz test Freemana-Haltona, odpowiednie do badania zależności w tabelach kontyngencji.

3.1 Zadanie 4

Zapoznaj się z funkcjami służącymi do wykonania testu Fishera oraz testu Freemana-Haltona.

```
# fisher.test(x, y = NULL, workspace = 200000, hybrid = FALSE,
#             hybridPars = c(expect = 5, percent = 80, Emin = 1),
#             control = list(), or = 1, alternative = "two.sided",
#             conf.int = TRUE, conf.level = 0.95,
#             simulate.p.value = FALSE, B = 2000)
```

Funkcja przyjmuje wiele argumentów, niektóre tylko w przypadku macierzy $ 2 ×2$. W formie w której zostało wyświetlone, najważniejsze parametry to:

- `x` tabela dwurymiarowa w formie macierzy lub typu `factor`
- `alternative` określa hipotezę alernatywną (H_1)
- `simulate.p.value` - wartość określająca sposób obliczania p-wartości w tablicach większych niż $ 2 ×2$

Przykładowe użycie funkcji - czy odpowiedzi na 2 pytania typu Tak/Nie są zależne?

```
# test fishera
odpowiedzi <-
matrix(c(13, 11, 5, 31),
       nrow = 2,
       dimnames = list(PYT1 = c("Tak", "Nie"),
                       PYT2 = c("Tak", "Nie")))
fisher.test(odpowiedzi)$p.value
```

```
[1] 0.001385665
```

P-value mniejsze od poziomu istotności, więc odrzucamy hipotezę H_0 o niezależności zmiennych. Wykonajmy również test dla tabeli o większych wymiarach, np. 2×3

```
# test freemana-haltona
odpowiedzi2 <- matrix(c(5, 10, 8, 5, 7, 8), nrow = 2, byrow = TRUE, dimnames = list(
  PYT1 = c("Tak", "Nie"), PYT2 = c("A", "B", "C")))
fisher.test(odpowiedzi2, simulate.p.value = TRUE, B = 100000)
```

```
    Fisher's Exact Test for Count Data with simulated p-value (based on
    1e+05 replicates)

data:  odpowiedzi2
p-value = 0.8604
alternative hypothesis: two.sided
```

P-value jest większe od poziomu istotności (przyjmujemy $\alpha = 0.05$), więc nie mamy podstaw aby odrzucić hipotezę H_0 o niezależności zmiennych.

3.2 Zadanie 5

Korzystajac z testu Fishera, na poziomie istotnosci 0.05, zweryfikuj hipotezę, że zmienna PŁEĆ i zmienna CZY_KIER są niezależne. Czy na poziomie istotności 0.05 możemy wnioskować, że prawdopodobieństwo tego, że na stanowisku kierowniczym pracuje kobieta jest równe prawdopodobieństwu tego, że na stanowisku kierowniczym pracuje mężczyzna? Uzasadnij odpowiedź.

```
[1] 0.6659029
```

Naszą hipotezą H_0 jest niezależność zmiennych. W naszym zadaniu zmienne, których niezależność testujemy to PŁEĆ oraz CZY_KIER. P-value jest wyższa od przyjętego poziomu ufności ($\alpha = 0.05$), więc nie ma podstaw do odrzucenia naszej hipotezy zerowej. W statystyce jednak nie mówimy, że "prawdopodobieństwo tego, że na stanowisku kierowniczym pracuje kobieta jest równe prawdopodobieństwu tego, że na stanowisku kierowniczym pracuje mężczyzna", a zamiast tego mówimy, że nie ma podstaw do odrzucenia tej hipotezy.

3.3 Zadanie 6

4 Część III

4.1 Zadanie 7 i 8

Zapoznaj się z funkcją służącą do wykonania testu niezależności chi-kwadrat i zweryfikuj hipotezę, że stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie nie zależy od zajmowanego stanowiska. Przyjmij poziom istotności 0.01. Wynik testu porównaj z wynikiem uzyskanym w zadaniu 6. Zaprezentuj reszty wyznaczane w teście na wykresie asocjacyjnym i dokonaj jego interpretacji.

```
dane <- matrix(c(20, 30, 25, 25), nrow = 2, byrow = TRUE)
colnames(dane) <- c("TAK", "NIE")
rownames(dane) <- c("TAK", "NIE")
dane
```

```
    TAK NIE
TAK  20  30
NIE  25  25
```

```r
test <- chisq.test(dane)
print(test)
```

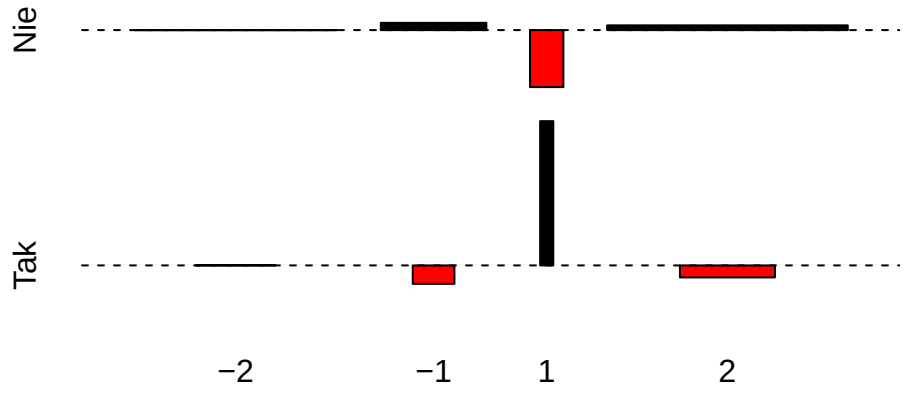
```
    Pearson's Chi-squared test with Yates' continuity correction

data:  dane
X-squared = 0.64646, df = 1, p-value = 0.4214
```

P-value wyszło większe niż 0.01, więc nie mamy podstaw do odrzucenia hipotezy zerowej. Zmienne są niezależne.

```
    Pearson's Chi-squared test

data:  tabela
X-squared = 13.114, df = 3, p-value = 0.004397
```



Czerwony słupek to reszta istotna. Czarny to reszta mało istotna.

4.2 Zadanie 9

Zapoznaj sie z funkcja słuzaca do generowania realizacji wektorów losowych z rozkładu wielomianowego, a nastepnie korzystajac z niej przeprowadz symulacje w celu oszacowania mocy testu Fishera oraz mocy testu chi-kwadrat Pearsona, generujac dane z tabeli 2×2, w której p11 = 1/40, p12 = 3/40, p21 = 19/40, p22 = 17/40. Symulacje wykonaj dla n =50, n=100 oraz n=1000. Sformułuj wnioski

```
         n=50 n=100 n=1000
pearson 0.218 0.360      1
fisher  0.038 0.226      1
```

Wnioski Dla większych n moc testu drastycznie rośnie.

4.3 Zadanie 10

Napisz funkcje, która dla danych z tablicy dwudzielczej oblicza wartosc poziomu krytycznego w tescie niezaleznosci opartym na ilorazie wiarogodnosci. Korzystajac z napisanej funkcji, wykonaj test dla danych przeanalizowanych w zadaniu 8.

```r
poziom_kretyczny <- function(zmienna1, zmienna2){
  tabela <- table(zmienna1, zmienna2)
  n_j <- colSums(tabela)
  n_i <- rowSums(tabela)
  n <- sum(n_i)
  lambda <- 1
  for (i in 1:nrow(tabela)) {
    for (j in 1:ncol(tabela)) {
      frac <- (n_i[i] * n_j[j])/ (tabela[i,j] * n)
      lambda <- lambda * frac^tabela[i, j]
    }

  }
  G_2 <- -2*log(lambda)
  p <- 1 - pchisq(G_2, (nrow(tabela)-1)*(ncol(tabela)-1) )
  return(p)
}
poziom_kretyczny(ankieta$PYT_2, ankieta$CZY_KIER)
```

```
       -2
0.03968956
```

5 Część IV i V

5.1 Zadanie 11

Przeprowadzone wsród brytyjskich mężczyzn badanie trwające 20 lat wykazało, że odsetek zmarłych (na rok) z powodu raka płuc wynosił 0,00140 wsród osób palących papierosy i 0,00010 wsród osób niepalących. Odsetek zmarłych z powodu choroby niedokrwiennej serca wynosił 0,00669 dla palaczy i 0,00413 dla osób niepalących. Opisz związek pomiędzy paleniem papierosów a śmiercią z powodu raka płuc oraz związek pomiędzy paleniem papierosów a śmiercią z powodu choroby serca. Skorzystaj z różnicy proporcji, ryzyka względnego i ilorazu szans. Zinterpretuj wartości. Związek której pary zmiennych jest silniejszy?

```
          Płuca   Serce
Pali     0.0014 0.00669
Nie pali 0.0001 0.00413
```

```
[1] "iloraz szans"
```

```
      płuca    serce
RP  0.00130 0.002560
RR 14.00000 1.619855
OR 14.01823 1.624029
```

WNIOSKI:

- Różnica proporcji w obu przypadkach jest niewielka: 0.0013 (płuca) oraz 0.00256 (serce). Wynika to z faktu, że podane prawdopodobieństwa były rzędu 0.001 lub mniejsze.
- Patrząc jednak na ryzyko względne (RR) widzimy, że w pierwszym przypadku wartość jest o wiele większa niż w drugim. Oznacza to, że dla osób cierpiących na raka płuc odsetek zmarłych był 14-krotnie większy dla osób palących niż niepalących. Dla chorych na serce różnica wynosi zdecydowanie mniej, około 1.6, jednak i tu widzimy, że większy odsetek był w grupie palących.
- Iloraz szans (OR) mówi nam, że szansa śmierci na raka płuc w grupie palących jest 14 razy większa niż w grupie niepalących a szansa śmierci na chorobę niedokrwienną serca około 1.6 razy większa dla palaczy niż niepalących.

5.2 Zadanie 12

Tabela 1 przedstawia wyniki dotyczące śmiertelności kierowców i pasażerów w wypadkach samochodowych na Florydzie w 2008 roku, w zależnożci od tego, czy osoba miała zapięty pas bezpieczeństwa czy nie.

Tabela 1	Śmiertelny	Nieśmiertelny
Bez pasów	1085	55 623
Z pasami	703	441 239

5.2.1 Zadanie 12.1

Oszacuj warunkowe prawdopodobieństwo śmierci w wypadku ze względu na drugą zmienną, tj. dla kierowców i pasażerów, który użyli pasa bezpieczeństwa oraz dla kierowców i pasażerów, który nie użyli pasa bezpieczeństwa.

```
     Śmiertelny
Pasy   tak    nie
  bez 1085  55623
  z    703 441239
```

```
  bez pasów    z pasami
0.019133103 0.001590706
```

5.2.2 Zadanie 12.2

Oszacuj warunkowe prawdopodobieństwo użycia pasa bezpieczeństwa ze względu na drugą zmienną, tj. dla kierowców i pasażerów ze śmiertelnymi obrażeniami oraz dla kierowców i pasażerów, którzy przeżyli wypadek.

```
   śmiertelne nieśmiertelne
    0.3931767     0.8880514
```

5.2.3 Zadanie 12.3

Jaki jest najbardziej naturalny wybór dla zmiennej objaśnianej w tym badaniu? Dla takiego wyboru wyznacz i zinterpretuj różnicę proporcji, ryzyko względne oraz iloraz szans. Dlaczego wartości ryzyka względnego i ilorazu szans przyjmują zbliżone wartości?

5.3 Zadanie 13

Oblicz wartości odpowiednich miar współzmienności (współczynnik tau lub współczynnik gamma) dla zmiennych:

5.3.1 Zadanie 13.1

stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie i zajmowane stanowisko,

```
[1] 0.0004091802
```

Bardzo małe τ oznacza bardzo słabą zależność (możemy przyjąć, że zmienne są niezależne).

5.3.2 Zadanie 13.2

stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie i staż pracy,

```
[1] 0.008886788
```

```
[1] 0.1435986
```

$\tau \approx 0.009$ oraz $\gamma > 0$ sugeruje, że mamy do czynienia z bardzo słabą dodatnią zależnością.

5.3.3 Zadanie 13.3

zajmowane stanowisko i staż pracy.

```
[1] 0.1158995
```

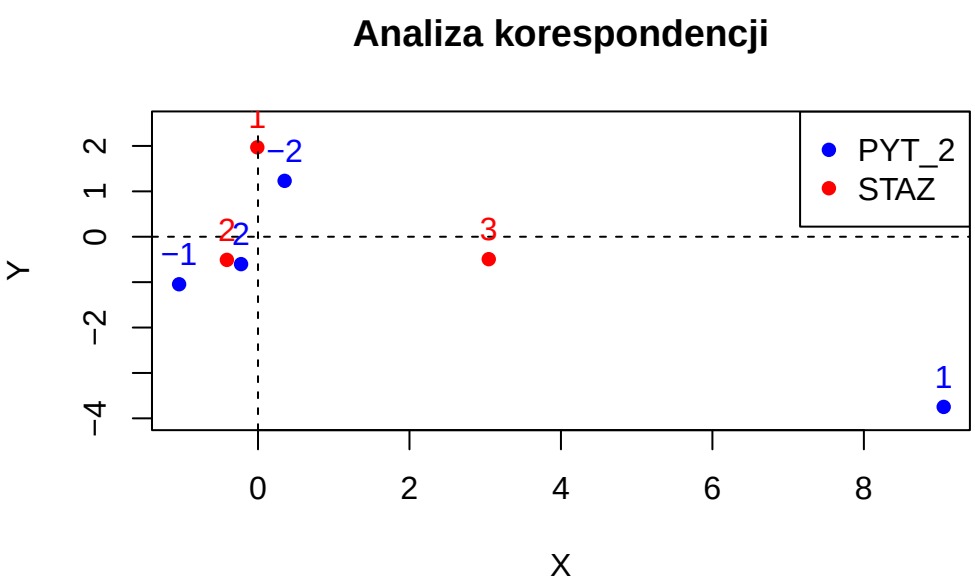
Dosyć małe τ oznacza słabą zależność zmiennych.

5.4 Zadanie 14

Na podstawie informacji przedstawionych na wykładzie napisz własną funkcję do przeprowadzania analizy korespondencji. Funkcja powinna przyjmować jako argument tablicę dwudzielczą i zwracać obliczone wartości odpowiednich wektorów i macierzy, współrzędnych punktów oraz odpowiedni wykres. Korzystając z napisanej funkcji wykonaj analizę korespondencji dla danych dotyczących stopnia zadowolenia ze szkoleń w kontekśie dopasowania do indywidualnych potrzeb w pierwszym badanym okresie oraz stażu pracy.

```
analiza_korespondencji <- function(zmienna1, zmienna2){
  tabela <- table(zmienna1, zmienna2)
  P <- as.matrix(tabela/sum(tabela))
  r <- rowSums(P)
  c <- colSums(P)
  Dr <- diag(r)
  Dc <- diag(c)
  Dr_1 <- solve(Dr)
  Dc_1 <- solve(Dc)
  R <- Dr_1 %*% P
  C <- P %*% Dc_1
  A <- Dr_1^(1/2) %*% (P - r %*% t(c)) %*% Dc_1^(1/2)
  b <- svd(A)
  U <- b$u
  V <- b$v
  F <- Dr_1^(1/2) %*% U
  G <- Dc_1^(1/2) %*% V
  row_stdx <- F[,1]
  row_stdy <- F[,2]
  col_stdx <- G[,1]
  col_stdy <- G[,2]

  plot(row_stdx, row_stdy, col = "blue", pch = 16, xlab = "X",
       ylab = "Y", main = "Analiza korespondencji",
       ylim = c(-4, 2.5), xlim=c(-1,9))
  points(col_stdx, col_stdy, col = "red", pch = 16)
  legend("topright", legend = c("PYT_2", "STAŻ"), col = c("blue", "red"), pch = 16)
  abline(h = 0, col = "black", lty = 2)
  abline(v = 0, col = "black", lty = 2)
  text(row_stdx, row_stdy, labels = as.character(c(-2,-1,1,2)), pos = 3, col = "blue")
  text(col_stdx, col_stdy, labels = as.character(c(1,2,3)), pos = 3, col = "red")
}
analiza_korespondencji(ankieta$PYT_2, ankieta$STAŻ)
```


Analiza korespondencji
PYT_2
STAZ
1
−2
2
2
−1
3
1
Y
X
−4
−2
0
1
2
0
2
4
6
8