Analiza danych ankietowych

Sprawozdanie 2

Weronika Jaszkiewicz Weronika Pyrtak

Spis treści

Część I

Zadanie 1

W firmie technologicznej przeprowadzono ankietę, w której 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 wyznaczono przedział ufności dla wektora prawodobieństw opisującego stopień zadowolenia ze szkolenia. Wybrano dwie metody dokładną Cloppera-Pearsona oraz asymptotyczną Wilsona. Przyjęto poziom ufności 0.95.

Tabela 1: Przedziały ufności dla dwóch estymacji

x	**Pierwsza estymacja**		**Druga estymacja**	
	Lower	**Upper**	**Lower**	**Upper**
14	0,0317	0,1299	0,0360	0,1316
17	0,0421	0,1487	0,0466	0,1500
40	0,1326	0,2822	0,1373	0,2819
100	0,4074	0,5926	0,4104	0,5896
29	0,0875	0,2200	0,0923	0,2205

Porównując pierwszą i drugą estymację, widać że przedziały ufności dla wszystkich wartości x są bardzo podobne, z minimalnymi przesunięciami w kierunku wyższych wartości w drugiej estymacji. Różnice są jednak bardzo niewielkie, co świadczy o stabilności oszacowań i sugeruje, że wyniki obu metod są zgodne.
W obu przypadkach szerokość przedziałów maleje wraz ze wzrostem liczby obserwacji x, co jest zgodne z intuicją — większe próby dają dokładniejsze oszacowania.

Zadanie 2

Napisano funkcję *testuj_hipoteze_multinomial()*, która wyznacza wartość poziomu krytycznego w następujących testach:

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

służacych do weryfikacji hipotezy $H_0 : p = p_0$ przy hipotezie alternatywnej $H_0 : p \neq p_0$ na podstawie obserwacji x wektora losowego X z rozkładu wielomianowego z parametrami n i p.
Funkcja przyjmuje dwa parametry:

- x - wektor obserwacji,
- p_0 - wektor hipotetycznych prawdopodobieństw.

Funkcja zwraca tabelę z wynikami testu: statystykę oraz p-value.

```r
testuj_hipoteze_multinomial <- function(x, p0) {
  # Dane wejściowe:
  # x  - wektor obserwacji (liczności)
  # p0 - wektor hipotetycznych prawdopodobieństw
  n <- sum(x)
  k <- length(x)
  expected <- n * p0

  chisq_stat <- sum((x - expected)^2 / expected)
  pval_chisq <- 1 - pchisq(chisq_stat, df = k - 1)

  nonzero <- x > 0
  g2_stat <- 2 * sum(x[nonzero] * log(x[nonzero] / expected[nonzero]))
  pval_g2 <- 1 - pchisq(g2_stat, df = k - 1)

  result <- data.frame(
    Test = c("Chi-kwadrat Pearsona",
             "Chi-kwadrat największej wiarygodności"),
    Statystyka = round(c(chisq_stat, g2_stat), 4),
    P_value = round(c(pval_chisq, pval_g2), 4)
  )
  return(result)
}
x <- c(14, 17, 40, 100, 29)
p0 <- rep(0.2, 5)
testuj_hipoteze_multinomial(x, p0)
```

Tabela 2: Wyniki testów chi-kwadrat dla weryfikacji hipotezy $H_0 : \mathbf{p} = \mathbf{p_0}$

Test	**Statystyka**	**P-value**
Chi-kwadrat Pearsona	123,1500	0,0000
Chi-kwadrat największej wiarygodności	106,1186	0,0000

Zarówno test chi-kwadrat Pearsona, jak i test chi-kwadrat największej wiarygodności dały bardzo wysokie wartości statystyk testowych (odpowiednio 123.15 oraz 106.12) oraz p-value

równe 0.
Przy standardowym poziomie istotności $\alpha = 0.05$ oba testy prowadzą do odrzucenia hipotezy zerowej $H_0 : p = p_0$. Oznacza to, że rozkład empiryczny obserwacji istotnie różni się od rozkładu teoretycznego zakładanego w hipotezie zerowej. Oba testy dają zgodne wnioski.

Zadanie 3

Na podstawie danych z ankiety z poprzedniej listy zweryfikowano 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", "zgadzam się", "zdecydowanie się zgadzam" na pytanie PYT_1. Przyjęto poziom istotności 0.05.

Tabela 3: Wyniki testów chi-kwadrat dla odpowiedzi na pytanie PYT_1 w Dziale Produkcyjnym

Test	**Statystyka**	**P-value**
Chi-kwadrat Pearsona	64,8571	0,0000
Chi-kwadrat największej wiarygodności	52,5271	0,0000

Przeprowadzone testy chi-kwadrat Pearsona oraz największej wiarygodności wskazują na bardzo wysokie wartości statystyk testowych (odpowiednio 64,86 oraz 52,53) oraz p-value równe 0. Przy przyjętym poziomie istotności $\alpha = 0.05$, w obu testach odrzucamy hipotezę zerową o równomiernym rozkładzie odpowiedzi na pytanie PYT_1 wśród pracowników Działu Produkcyjnego.

Oznacza to, że rozkład odpowiedzi na PYT_1 nie jest równomierny — nie wszystkie odpowiedzi ("zdecydowanie się nie zgadzam", "nie zgadzam się", "nie mam zdania", "zgadzam się", "zdecydowanie się zgadzam") są jednakowo prawdopodobne.

Część II

Zadanie 4

Test Fishera, znany również jako **test F** lub **test jednorodności wariancji**, jest statystycznym testem używanym do porównywania wariancji dwóch próbek populacji. Służy do określenia, czy wariancje dwóch grup są równe czy też nie. Test opiera się na rozkładzie F, który jest rozkładem prawdopodobieństwa porównującym stosunek dwóch wariancji. Jeśli hipoteza zerowa jest prawdziwa, stosunek wariancji powinien być bliski 1. Jeśli stosunek jest znacznie różny od 1, sugeruje to, że wariancje dwóch grup nie są równe.

Test Fishera dla tabel $r \times c$ zwany jest również **testem Freemana-Haltona**. Test ten jest rozszerzeniem na tabele $r \times c$ dokaładnego testu Fishera. Określa dokładne prawdopodobieństwo wystąpienia konkretnego rozkładu liczb w tabeli przy znanym n i ustalonych sumach brzegowych.

W języku programowania R w paczce *stats* występuje funkcja *fisher.test()* służąca do przeprowadzania testu Fishera dla przypadku 2×2 jak również przypadków o większej liczbie kolumn/wierszy. Funkacj przyjmuje parametry:

x – macierz 2-wymiarowa (tabela kontyngencji) lub obiekt typu faktor.

y – drugi faktor (używany tylko, jeśli x nie jest macierzą).

workspace – liczba całkowita określająca ilość pamięci do obliczeń dla dużych tabel (w jednostkach po 4 bajty).

hybrid – czy dla dużych tabel stosować przybliżoną metodę hybrydową (domyślnie: $FALSE$, czyli dokładnie).

hybridPars – parametry sterujące warunkami przybliżenia chi-kwadrat (domyślnie tzw. warunki Cochrana).

control – lista ustawień technicznych dla algorytmu (np. *mult* – rozmiar pamięci na ścieżki w obliczeniach).

or – zakładany iloraz szans (odds ratio), używany tylko w przypadku tabel 2×2.

alternative – rodzaj hipotezy alternatywnej: *"two.sided"* (dwustronna), *"greater"* lub *"less"*.

conf.int – czy obliczyć przedział ufności dla ilorazu szans (tylko dla 2×2).

conf.level – poziom ufności dla przedziału (np. 0.95 dla 95%).

simulate.p.value – czy obliczyć p-value przy pomocy symulacji Monte Carlo (zalecane dla dużych tabel).

B – liczba powtórzeń w symulacji Monte Carlo.

Funkcja zwraca:

method – nazwa zastosowanego testu, np. "Fisher's Exact Test for Count Data"

data.name – nazwa tabeli kontyngencji przekazanej do testu

p.value – wartość p, określająca istotność statystyczną wyniku

alternative – określenie hipotezy alternatywnej: "two.sided" (dwustronna), "less" lub "greater"

estimate – oszacowany iloraz szans (odds ratio), zwracany tylko w przypadku tabeli 2×2

conf.int – przedział ufności dla ilorazu szans, domyślnie 95% (również tylko dla tabeli 2×2)

null.value – wartość zakładana w hipotezie zerowej (np. iloraz szans = 1)

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x_freeman_halton
## p-value = 1
## alternative hypothesis: two.sided
```

Wyniki testu Fishera dla danych liczbowych:

1. **Test dla x_freeman_halton:**

Wynik testu wskazuje, że nie ma dowodów na to, by istnieła istotna statystycznie różnica między grupami. Przy p-value równym 1, nie możemy odrzucić hipotezy zerowej (brak związku).

Hipoteza alternatywna: Dwustronna (testuje, czy iloraz szans jest różny od 1).

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x_fisher
## p-value = 0.03497
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##     1.008849 1049.791446
## sample estimates:
## odds ratio
##   15.46969
```

2. **Test dla x_fisher:**

Wynik wskazuje na statystycznie istotną różnicę, ponieważ p-value (0.03497) jest mniejsze niż standardowy próg 0.05. Oznacza to, że istnieje związek między badanymi zmiennymi.

Hipoteza alternatywna: Zakłada, że iloraz szans nie jest równy 1 (testuje, czy szanse w grupach się różnią).

Przedział ufności (95%) dla ilorazu szans: [1.008849, 1049.791446], co oznacza, że z 95% pewnością, iloraz szans w populacji mieści się w tym przedziale.

Iloraz szans (odds ratio): 15.46969 – sugeruje, że jedna z grup ma około 15.47 razy większe szanse na daną cechę w porównaniu do drugiej grupy.

Zadanie 5

Wkorzystując test Fishera, na poziomie istotnosci 0.05, zweryfikowano hipotezę, że nie istnieje zależność między płcią a zajmowaniem kierowniczego stanowiska.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 0.6659
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5299411 3.8023038
## sample estimates:
## odds ratio
##   1.358208
```

Na podstawie analizy i przy poziomie istotności 0.05, nie możemy stwierdzić, że kobiety i mężczyźni różnią się pod względem prawdopodobieństwa bycia kierownikiem – uznajemy, że prawdopodobieństwa są równe.

Zadanie 6

Wykorzystując test Freemana-Haltona na poziomie istotnosci 0.05 zweryfikowano następujące hipotezy:

a) Zajmowanie stanowiska kierowniczego nie zależy od wieku.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 0.7823
## alternative hypothesis: two.sided
```

Na podstawie analizy, nie możemy stwierdzić, że wiek wpływa na prawdopodobieństwo bycia kierownikiem – uznajemy, że prawdopodobieństwa są równe.

b) Zajmowanie stanowiska kierowniczego nie zależy od stażu pracy.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 6.538e-05
```

```
## alternative hypothesis: two.sided
```

Na podstawie analizy, możemy stwierdzić, że jest powiązanie między stażem pracy a zajmowaniem stanowiska kierowniczego.

c) Stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie nie zależy od zajmowanego stanowiska.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 0.0443
## alternative hypothesis: two.sided
```

Na podstawie analizy, można zatem stwierdzić, że poziom zadowowolenia z dopasowania szkoleń do potrzeb zależy od zajmowanego stanowiska w pierwszym badanym okresie.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 0.8377
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4612836 2.8018002
## sample estimates:
## odds ratio
##   1.125705
```

Na podstawie wyników testu nie można stwierdzić istotnej zależności między zajmowaniem stanowiska kierowniczego a zadowoleniem ze szkoleń.

d) Stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie nie zależy od stażu.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 0.01069
## alternative hypothesis: two.sided
```

Na podstawie analizy, można zatem stwierdzić, że poziom zadowolenia z dopasowania szkoleń do potrzeb zależy od stażu pracy w pierwszym badanym okresie.

```
##
##  Fisher's Exact Test for Count Data
##
```

```
## data:  tabela
## p-value = 0.4097
## alternative hypothesis: two.sided
```

Na podstawie wyników testu nie można stwierdzić istotnej zależności między stażem pracy a zadowoleniem ze szkoleń.

e) Stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie nie zale˙zy od płci.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 0.4758
## alternative hypothesis: two.sided
```

Na podstawie analizy, nie można stwierdzić zależności poziomu zadowolenia z dopasowania szkoleń do potrzeb względem płci w pierwszym badanym okresie.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 0.6589
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.6194413 2.1460710
## sample estimates:
## odds ratio
##   1.152656
```

Na podstawie wyników testu nie można stwierdzić istotnej zależności między płcią pracownika a zadowoleniem ze szkoleń.

f) Stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie nie zależy od wieku.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabela
## p-value = 0.3194
## alternative hypothesis: two.sided
```

Na podstawie wyników testu, nie można stwierdzić zależności poziomem zadowolenia dopasowania szkoleń do potrzeb względem płci w pierwszym badanym okresie.

```
##
##  Fisher's Exact Test for Count Data
```

```
##
## data:  tabela
## p-value = 0.3275
## alternative hypothesis: two.sided
```

Na podstawie wyników testu nie można stwierdzić istotnej zależności między płcią pracownika a zadowoleniem ze szkoleń.

Podsumowując, można zauważyć, że wynik testu zależy od tego, czy analizujemy pełną skalę ocen zadowolenia ze szkoleń (od -2 do 2), czy też sprowadzamy odpowiedzi do postaci binarnej (TAK/NIE). Można zauważyć, że bardziej szczegółowe dane (skala porządkowa) mogą dostarczyć dokładniejszych i bardziej czułych wyników analizy statystycznej niż uproszczona zmienna binarna, która może zatracać istotne różnice w ocenach.

Część III

Zadanie 7

Funkcja *chisq.test()*, służy do wykonania testu niezależności chi-kwadrat. Przyjmuje argumenty:

- x: Wektor liczbowy, macierz lub czynnik,
- y: Drugi wektor (ignorowany, jeśli x to macierz), jeśli oba są czynnikami.
- correct: Czy stosować korektę ciągłości (dla tabel 2x2).
- p: Wektor prawdopodobieństw,
- rescale.p: Czy skalować p, by sumowało się do 1,
- simulate.p.value: Czy obliczać p wartość metodą Monte Carlo.
- B: Liczba replikacji Monte Carlo.

Zwraca ona:

- statystykę testu chi-kwadrat,
- stopnie swobody,
- p-value.

Na tej podstawie można ocenić, czy zmienne są niezależne.

Hipoteza zerowa H_0: zmienne są niezależne, hipoteza alternatywna H_1 : zmienna są zależne.

```
tablica <- matrix(c(10, 20, 30, 40), nrow = 2, byrow = TRUE)
chisq.test(tablica)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tablica
## X-squared = 0.44643, df = 1, p-value = 0.504
```

Biorąc poziom istotności $\alpha = 0.05$, nie ma podstaw do odrzucenia hipotezy zerowej o niezależności zmiennych. Oznacza to, że brak jest statystycznie istotnych dowodów na istnienie zależności pomiędzy badanymi cechami.

Zadanie 8

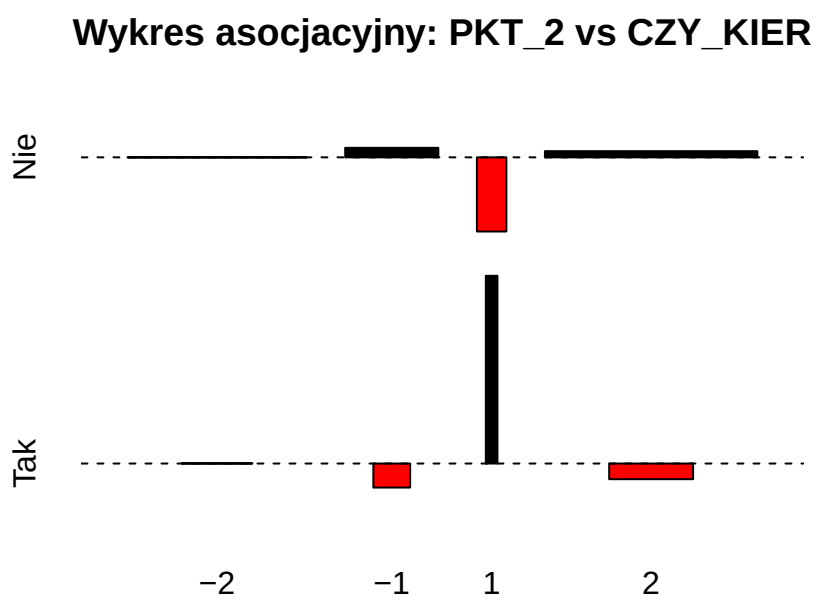
Korzystając z funkcji poznanej w zadaniu 7. zweryfikowano hipotezę, że stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie nie zależy od zajmowanego stanowiska. Przyjęto poziom istotności 0.01.

Zatem hipotezą zerową H_0 jest: PYT_2 i CZY_KIER są niezależne.

Hipotezą alternatywną H_1 jest: PYT_2 i CZY_KIER są zależne.

Tabela 4: Reszty standaryzowane z testu chi-kwadrat Pearsona

PYT_2	**Nie**	**Tak**
-2	-0,0043	0,0043
-1	0,4828	-0,4828
1	-3,5978	3,5978
2	0,4307	-0,4307



Stąd

$$p_{value} = 0.004397 < \alpha = 0.01$$

Zatem odrzucamy hipotezę zerową. Wyniki testu chi-kwadrat wskazują na statystycznie istotną zależność między oceną dopasowania szkoleń a tym, czy ktoś pełni funkcję kierowniczą. Przedstawiono również reszty standaryzowane. Reszty są różnicą miezy wartościami obserwowanymi a oczekiwanymi. Jednak różnica standaryzowana jest dana wzorem:

$$R = \frac{O - E}{\sqrt{E}}$$

gdzie:

- O - wartość obserwowana,
- E - wartość oczekiwana.

Test chi-kwadrat mówi czy jest zależność, a reszty pokazują gdzie dokładnie ona jest. Reszty pokazują, które kombinacje zmiennych łamią założenie niezależności. Gdy $R = 0$, to wartość obserwowana i oczekiwana są do siebie zbliżone, nie daje nam to efektu. Gdy

$|R| > 2$, różnica jest istotna statystycznie.

W powstałej tabeli reszt standaryzowanych, wartość $|R| > 2$ w PYT_2=1. Zatem kierownicy odpowiedzieli '1' o wiele więcej razy niż się oczekiwano, a nie-kierownicy odpowiedzieli '1' o wiele mniej razy niż się oczekiwano, co wskazuje na złamanie założenia niezależności. Pozostałe wartości reszt są niewielkie, zatem one nie łamią założenia niezależności.
Następnie przedstawiono wykres asocjacyjny, który ukazuje reszty standaryzowane dla tabeli kontyngencji. Każdy słupek odnosi się do jednej z kategorii PYT_2. Oś Y przedstawia odpowiedzi CZY_KIER - "TAK" lub "NIE".

Kolor czerwony słupka określa, że jest mniej przypadków niż oczekiwano (ujemna reszta), a czarny, że więcej przypadków niż oczekiwano (dodatnia reszta).Wartości z tabeli wyraźnie widać na wykresie - dla odpowiedzi '1' w PYT_2, pojawiają się największe reszty standaryzowane, a więc najsilniejsze odchylenia od niezależności.

Analiza wykresu asocjacyjnego wskazuje, że osoby na stanowiskach kierowniczych częściej oceniały szkolenia jako średnio dopasowane, natomiast osoby niepełniące funkcji kierowniczych częściej udzielały ocen skrajnych (niskich lub wysokich).

Wyniki z Zadania 8 są bardziej przekonujące dzięki analizie bardziej szczegółowych danych i użyciu testu chi-kwadrat. Pokazują, że zajmowanie stanowiska kierowniczego ma wpływ na ocenę dopasowania szkoleń. Zadanie 6 potwierdza to w ograniczonym zakresie – zależność była widoczna jedynie w jednej z wersji kodowania danych.

Zadanie 9

Przeprowadzono symulacje w celu oszacowania mocy testu Fishera oraz mocy testu chi--kwadrat Pearsona, generując dane z tabeli 2×2, w której:

- $p_{11} = 1/40$,
- $p_{12} = 3/40$,
- $p_{21} = 19/40$,
- $p_{22} = 17/40$.

500 symulacji wykonano dla $n = 50$, $n = 100$ oraz $n = 1000$.

Tabela 5: Porównanie mocy testów chi-kwadrat i Fishera w zależności od liczności próby

Liczność próby (n)	**Test chi-kwadrat**	**Test Fishera**
50	0,072	0,100
100	0,260	0,318
1000	1,000	0,998

Przeprowadzone symulacje pokazują, że moc testu statystycznego silnie zależy od liczności próby. Dla małych prób (n = 50) zarówno test chi-kwadrat, jak i test Fishera mają niską

moc, co może prowadzić do niewykrycia zależności w danych.Oznacza to dużą szansę na błąd II rodzaju, czyli nieodrzucenie fałszywej hipotezy zerowej. Dla n = 100 moc rośnie, ale nadal może być niewystarczająca w badaniach wymagających dużej czułości. Dopiero przy dużej liczności próby (n = 1000) oba testy niemal zawsze wykrywają zależność. Test Fishera okazuje się nieco bardziej efektywny przy małych próbach.

Zadanie 10

Napisano funkcję, która dla danych z tablicy dwudzielczej oblicza wartość poziomu krytycznego w teście niezależnosci opartym na ilorazie wiarogodności. Korzystając z napisanej funkcji, wykonano test dla danych przeanalizowanych w zadaniu 8.

Użyto poniższych wzorów:

- Decyzja (iloraz wiarygodności):

$$\lambda = \prod_{i,j} \left(\frac{n_{i+} \cdot n_{+j}}{n \cdot n_{ij}} \right)^{n_{ij}}$$

- Statystyka testowa:

$$G^2 = -2 \log \lambda$$

- Wartość krytyczna (p-value):

$$p = 1 - F_{\chi^2_{(r-1)(c-1)}}(G^2)$$

```
test_IW <- function(tabela) {
  n <- sum(tabela)
  wiersze <- rowSums(tabela)
  kolumny <- colSums(tabela)

  E <- outer(wiersze, kolumny, FUN = function(a, b) a * b / n)
  G2 <- 2 * sum(tabela * log(tabela / E), na.rm = TRUE)

  df <- (nrow(tabela) - 1) * (ncol(tabela) - 1)
  p_value <- 1 - pchisq(G2, df)

  return(list(G2 = G2, df = df, p_value = p_value))
}
tabela <- table(df$PYT_2, df$CZY_KIER)
test_IW(tabela)
```

Otrzymana wartość statystyki wyniosła $G^2 = 8.33$ przy 3 stopniach swobody, a odpowiadająca jej wartość p wyniosła $p = 0.0397$. Przy założonym poziomie istotności $\alpha = 0.01$, nie ma podstaw do odrzucenia hipotezy zerowej o niezależności badanych zmiennych ($p > \alpha$).

Oznacza to, że test G^2 nie wykazał istotnego statystycznie związku między oceną dopasowania szkoleń do indywidualnych potrzeb (PYT_2) a pełnieniem funkcji kierowniczej (CZY_KIER). Wynik ten jest nieco odmienny od wyniku testu chi-kwadrat z zadania 8, w którym zależność uznano za istotną. Może to wynikać z różnic w czułości testów przy danej liczności próby.

Część IV i V

Zadanie 11

Celem jest analiza związku między paleniem papierosów a ryzykiem śmiercie z powodu raka płuc oraz choroby niedokrwiennej serca. Na podstawie danych obliczono różnicę proporcji, ryzyko względne oraz iloraz szans, które zastosowano do oceny sił zależności.

Różnica proporcji (Risk Difference, RD):
Różnica między proporcją osób, które doświadczyły danego zdarzenia (np. śmierci) w dwóch grupach (np. palacze vs. niepalacze). Mierzy absolutną różnicę w ryzyku.

Ryzyko względne (Relative Risk, RR):
Stosunek ryzyka wystąpienia zdarzenia w jednej grupie do ryzyka w drugiej grupie. Mówi, jak wiele razy jedno ryzyko jest wyższe od drugiego.

Iloraz szans (Odds Ratio, OR):
Stosunek szans (odds) wystąpienia zdarzenia w jednej grupie do szans wystąpienia tego samego zdarzenia w drugiej grupie. Pomaga w ocenie siły związku między zmiennymi, szczególnie w badaniach przypadków-kontrolnych.

Choroba	Palacz	Niepalacz
Rak płuc	0.00140	0.00010
Niedokrwienie serca	0.00669	0.00413

Choroba	RP	RR	OR
Rak płuc	0.00130	14.00	14.02
Niedokrwienie serca	0.00256	1.62	1.62

Rak płuc

Różnica proporcji (RP = 0.00130) Oznacza, że w grupie palaczy umiera rocznie o 0.13% więcej osób z powodu raka płuc niż w grupie niepalących.

Ryzyko względne (RR = 14) Palacze mają 14 razy wyższe ryzyko śmierci z powodu raka płuc w porównaniu do osób niepalących. To bardzo silny związek.

Iloraz szans (OR = 14.018) Wskazuje, że szansa na śmierć z powodu raka płuc jest ponad 14 razy większa u palaczy niż u niepalących. Potwierdza to silną zależność.

Choroba niedokrwienna serca

Różnica proporcji (RP = 0.00256) Oznacza, że wśród palaczy rocznie umiera o 0.256% więcej osób z powodu choroby serca niż wśród niepalących — różnica większa bezwzględnie niż dla raka płuc, ale…

Ryzyko względne (RR 1.62) Palacze mają około 1.62 razy wyższe ryzyko śmierci z powodu choroby serca niż niepalący — związek jest obecny, ale znacznie słabszy niż w przypadku raka płuc.

Iloraz szans (OR 1.62) Szansa na śmierć z powodu choroby serca jest około 1.62 razy większa u palaczy niż u niepalących.

Podsumowując, choć różnica proporcji (RP) jest większa dla choroby serca, to ryzyko względne i iloraz szans — które lepiej mierzą siłę zależności — są znacznie większe dla raka płuc. To oznacza, że palenie papierosów ma znacznie silniejszy związek ze śmiercią z powodu raka płuc niż ze śmiercią z powodu choroby niedokrwiennej serca.

Zadanie 12

12.1

Wskaźnik	Wartość
Prawdopodobieństwo śmierci (pasażerowie bez pasów)	0.00159
Prawdopodobieństwo śmierci (pasażerowie z pasami)	0.01913

12.2

Wskaźnik	Wartość
Proporcja pasażerów z pasami wśród zmarłych	0.39318
Proporcja pasażerów z pasami wśród żywych	0.88805

12.3

Najbardziej naturalnym wyborem zmiennej objaśnianej w tym badaniu jest **śmierć w wypadku** (śmiertelny vs. niesmiertelny), ponieważ celem badania jest ocena, jak użycie pasów wpływa na ryzyko śmierci w wypadku.

Wskaźnik	Wartość
Różnica proporcji:	0.01754
Ryzyko względne:	12.02805
Iloraz szans:	12.24317

Różnica proporcji:

Różnica proporcji wynosi 1.75%,co oznacza, że brak pasów zwiększa ryzyko śmierci o 1,75 punktu procentowego.

Ryzyko względne (RR):

Ryzyko względne wynosi 12,03, co oznacza, że osoby bez pasów miały ponad 12 razy większe ryzyko śmierci niż osoby z pasami.

Iloraz szans :

Iloraz szans wynosi 12.2, co oznacza, że osoby, które nie używają pasów bezpieczeństwa, mają 12.2 razy wyższe szanse na śmierć w wypadku niż osoby, które używają pasów.

Ryzyko względne (RR) i iloraz szans (OR) przyjmują podobne wartości, ponieważ śmiertelność w analizowanych danych jest niewielka – czyli mamy do czynienia z rzadkim zdarzeniem. W takich sytuacjach:

$$p \ll 1 \Rightarrow p \approx \text{szansa (odds)}$$

Dlatego:

$$RR \approx OR$$

Zadanie 13

Miary współzmienności to liczby opisujące relacje między zmiennymi. Przykładowymi miarami współzmienności są *tau* i *gamma*.

Tau to miara współzależności zmiennych nominalnych, stosowana w tabelach krzyżowych. Opiera się na losowym przypisaniu kategorii i mierzy, o ile lepiej możemy przewidzieć wartość jednej zmiennej, znając wartość drugiej, w porównaniu do losowego zgadywania. Im wyższa wartość tau (maksymalnie 1), tym silniejszy związek między zmiennymi.

Gamma to miara współzależności zmiennych porządkowych. Mierzy siłę i kierunek związku między dwiema zmiennymi, uwzględniając jedynie ich rangi. Gamma ocenia, jak często zmiany w jednej zmiennej powodują zmiany w drugiej w tym samym lub przeciwnym kierunku. Wartość gamma mieści się w przedziale od -1 do 1: wartość 1 oznacza doskonałą dodatnią zależność, -1 doskonałą ujemną zależność, a 0 brak zależności.

1. Miara współzmienności dla stopnia zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie i zajmowanego stanowiska:

```
## [1] "Współczynnik tau wynosi: 0.00168"
```

Wartość współczynnika tau jest bardzo bliska zeru, co sugeruje, że między zmiennymi nie występuje istotny związek. Stopień zadowolenia ze szkoleń nie różnicuje się wyraźnie w zależności od zajmowanego stanowiska.

2. Miary współzmienności dla stopnia zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie i stażu pracy:

```
## [1] "Współczynnik tau wynosi: 0.0191"
```

```
## [1] "Współczynnik gamma wynosi: 0.09084"
```

Wartość współczynnika *tau* wskazuje, że znajomość poziomu zadowolenia ze szkoleń pozwala przewidzieć długość stażu pracy jedynie o około 2% lepiej niż bez tej wiedzy. Jest to wynik bardzo słaby, co sugeruje brak istotnej zależności między zmiennymi.

Wartość współczynnika *gamma* sugeruje słabą dodatnią zależność: wzrost jednego z parametrów wiąże się z niewielkim wzrostem drugiego w tym samym kierunku. Siła tej zależności jest jednak niska.

3. Miara współzmienności dla zajmowanego stanowiska i stażu pracy:

```
## [1] "Współczynnik tau wynosi: 0.1159"
```

Wartość miary tau oznacza, że znajomość jednej zmiennej pozwala przewidzieć drugą o około 11,6% lepiej niż przy losowym doborze. Jest to niewielka, lecz zauważalna zależność, co może sugerować, że osoby na wyższych stanowiskach mają tendencję do dłuższego stażu pracy, ale związek ten nie jest silny.

Zadanie 14

Analiza korespondencji to metoda eksploracyjna służąca do badania zależności między kategoriami dwóch zmiennych przedstawionych w tabeli kontyngencji.

```r
correspondence_analysis <- function(tabela){
  "Przedstawienie zależności między wierszami i kolumnami tabeli kontyngencji"

n <- sum(tabela)
p_matrix <- as.matrix(tabela/n) # normalizacja tabeli

#Wektory częstości brzegowych
r <- rowSums(p_matrix) # wierszy
c <- colSums(p_matrix) # kolumn
names(r) <- NULL
names(c) <- NULL

D_r <- diag(r^(-1/2)) # macierz częstości wierszowych (podniesiona do -1/2 potęgi)
D_c <- diag(c^(-1/2)) # macierz częsctości kolmnowych (podniesiona do -1/2 potęgi)

R <- solve(D_r) %*% p_matrix # macierz profili wierszowych (macierz częstości wierszow
C <- p_matrix %*% solve(D_c) # macierz profili kolumnowych (macierz częsctości kolmnow

A <- D_r%*%(p_matrix - r%*%t(c))%*%D_c # macierz residuów standaryzowanych

svd_result <- svd(A)
U <- svd_result$u
Gamma <- diag(svd_result$d)
V <- svd_result$v
```

```
F <- D_r%*%U%*%Gamma # współrzędne kategorii cech dla wierszy
G <- D_c%*%V%*%Gamma # współrzędne kategorii cech dla kolumn

row_x <- F[,1]
row_y <- F[,2]

col_x <- G[,1]
col_y <- G[,2]

# Wykres
xlim <- range(c(row_x, col_x))
ylim <- range(c(row_y, col_y))

plot(NA, NA, xlim = xlim, ylim = ylim,
     xlab = "Dimension 1", ylab = "Dimension 2",
     main = "Correspondence Analysis Plot", asp = 1)

points(row_x, row_y, col = "blue", pch = 1)
points(col_x, col_y, col = "red", pch = 17)
abline(h = 0, v = 0, lty = 2)
text(row_x, row_y, labels = rownames(tabela), pos = 3, col = "blue", cex = 0.8)
text(col_x, col_y, labels = colnames(tabela), pos = 3, col = "red", cex = 0.8)

return(list(
  mass_rows = r,
  mass_cols = c,
  row_coords = F,
  col_coords = G
))
}
```

Korzystając z powyższej funkcji przeprowadzono analizę korespondencji dla danych dotyczących stopnia zadowolenia ze szkoleń w konteście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie oraz stażu pracy. Tabela kontyngencji została utworzona na podstawie odpowiedzi na PYT_2 (wiersze) i zmiennej $STAŻ$ (kolumny). Celem analizy było zbadanie powiązań między typami odpowiedzi a grupami osób o różnym stażu.

Wyniki zostały przedstawione za pomocą odpowiednich macierzy oraz wykresu.

Analiza wykresu:

Niebieskie okręgi reprezentują poszczególne odpowiedzi na pytanie PYT_2 (czyli kategorie wierszy).

Czerwone trójkąty to grupy stażu ($STAŻ$) – kategorie kolumn.

Współrzędne punktów wyznaczane są w dwóch wymiarach – Dimension 1 i Dimension 2,

które pokazują główne osie zmienności w danych (czyli różnice między kategoriami).

Im dalej od środka (punktu przecięcia osi), tym bardziej dana kategoria odstaje od „przeciętnego profilu", czyli ma bardziej unikalny rozkład względem drugiej zmiennej.

Odpowiedź na PYT_2 oznaczone "1" znajduje się daleko po prawej stronie na osi Dimension 1, co oznacza, że ta odpowiedź znacznie różni się od innych – jej profil odpowiedzi jest specyficzny. Najprawdopodobniej występowała głównie w jednej grupie stażu, prawdopodobnie były to osoby ze stażem ponad trzyletnim, ponieważ punkty leżą blisko siebie.

Odpowiedzi "2", "-1" i "-2" Znajdują się po stronie lewej, w pobliżu środka. Są bliżej siebie, więc ich rozkład odpowiedzi był podobny. Leżą również relatywnie blisko punktu kolumny „2" oraz kolumny "1", co sugeruje, że były często wybierane przez osoby z tej grupy stażu.

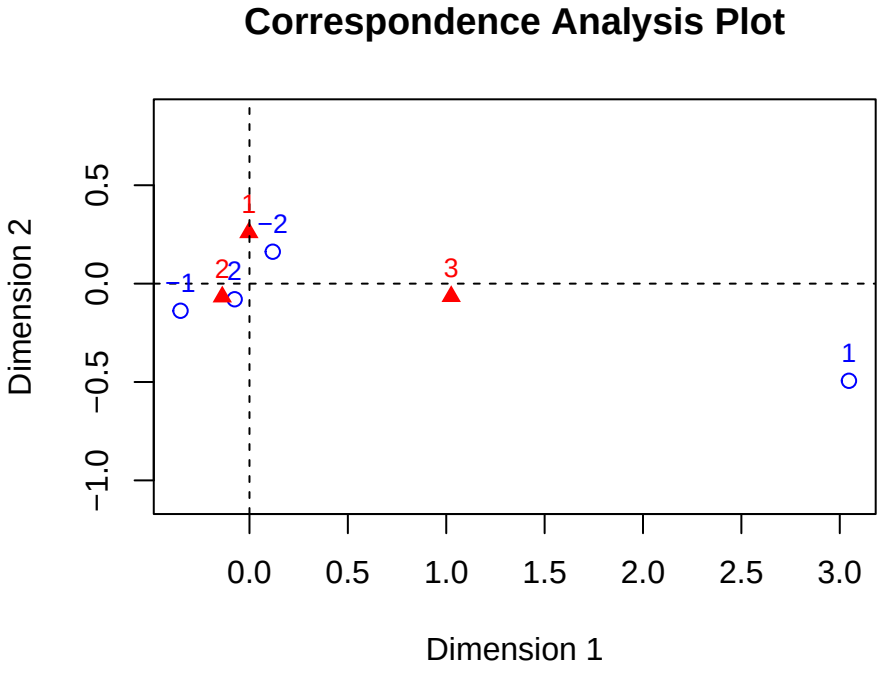
Intepretacja wyników:

mass_rows: W ankiecie najwięcej zostało udzielonych odpowiedzi 2 (52%) oraz -2 (37%), co oznacza, że to one najbardziej wpływają na strukturę danych. Natomiast odpowiedź 1 ma znikomy wpływ (1%)

mass_cols: Dominującym stażem wśró ankietowanych jest odpowiedź 2 (70%), czyli ta grupa badanych występuje najczęściej i w największym stopniu kształtuje zależności między zmiennymi. Najmniejszą grupą są osoby ze stażem 3 (9.5%).

row_coords: Odpowiedź -1 silnie odstaje w pierwszym wymiarze (3.05), co oznacza, że zachowuje się zupełnie inaczej niż pozostałe i najprawdopodobniej jest powiązana z jedną konkretną kategorią kolumn; pozostałe kategorie są bardziej zbliżone do centrum.

col_coords: Staż pracy powyżej 3 lat (1.02) wyróżnia się w pierwszym wymiarze, co sugeruje silny związek z wyjątkową kategorią wierszy (najpewniej tą, która odstaje), natomiast pozostałe są mniej zróżnicowane.



```
## $mass_rows
## [1] 0.37 0.10 0.01 0.52
##
## $mass_cols
## [1] 0.205 0.700 0.095
##
## $row_coords
##             [,1]       [,2]          [,3]
## [1,]  0.11826617  0.1622275 -4.508564e-17
## [2,] -0.35056182 -0.1378599  1.285171e-18
## [3,]  3.04672664 -0.4937333 -1.608678e-17
## [4,] -0.07532609 -0.0794247 -6.403064e-17
##
## $col_coords
##              [,1]        [,2]         [,3]
## [1,] -0.003060932  0.25943499 5.372911e-17
## [2,] -0.138246087 -0.06714975 5.372911e-17
## [3,]  1.025260549 -0.06504575 5.372911e-17
```

Część dodatkowa

Zadanie 1

Napisano funkcję, która dla dwóch wektorów danych oblicza wartość poziomu krytycznego (p-value) w teście opartym na korelacji odległości. Następnie dla wygenerowanych danych zweryfikowano hipotezę o niezależności przy użyciu napisanej funkcji.
Pod uwagę wzięto dwa przypadki. W pierwszym przykładzie wektory X i Y są niezależne, gdzie $X_n, Y_n \sim N(0,1)$. W drugim wektory są zależne: $X_n \sim N(0,1)$ oraz $Y_n \sim X_n^2 + N(0,0.1)$. Za poziom istotności przyjęto 0.05.

```r
library(energy)

test_korelacji_odleglosci <- function(x, y, R = 499) {
  wynik <- dcor.test(x, y, R = R)
  return(wynik$p.value)
}
set.seed(123)
x <- rnorm(100)
y <- rnorm(100)

test_korelacji_odleglosci(x, y)
# p > 0.05 → nie odrzucamy H  (brak zależności)
x <- rnorm(100)
y <- x^2 + rnorm(100, 0, 0.1)

test_korelacji_odleglosci(x, y)
# p < 0.05 → odrzucamy H , zależność wykryta!
```

Dla danych niezależnych mamy $p = 0.656$, czyli nie ma podstaw do odrzucenia hipotezy zerowej o niezależności zmiennych. Oznacza to, że test nie wykrył zależności między X i Y, co jest zgodne z założeniem, że dane są niezależne.
Dla drugiego przypadku $p = 0.002$, czyli $p < 0.05$ Hipoteza zerowa została odrzucona — test wykrył istotną zależność między X a Y. Co ważne, test oparty na korelacji odległościowej potrafi wykrywać również nieliniowe zależności, więc nawet jeśli korelacja liniowa byłaby bliska zeru, związek może być obecny.

Zadanie 2

Dla zadanych π_1 oraz π_2 pokazano, że wartość ryzyka względnego (RR) nie jest bardziej oddalona od wartości 1 (wartość odpowiadająca niezależności) niż wartość odpowiadającego ilorazu szans (OR).
Innymi słowy należy pokazać, że:

$$|RR - 1| \leq |OR - 1|$$

dla dwóch prawdopodobieństw:

- $\pi_1 = P(Zdarzenie|Grupa1)$,
- $\pi_2 = P(Zdarzenie|Grupa2)$.

gdzie $\pi_1, \pi_2 \in (0, 1)$.

Obliczamy ryzyko względne:

$$RR = \frac{\pi_1}{\pi_2}$$

oraz iloraz szans:

$$OR = \frac{\pi_1(1-\pi_2)}{\pi_2(1-pi_1)}$$

Wtedy

$$L = \left|\frac{\pi_1}{\pi_2} - 1\right| = \left|\frac{\pi_1 - \pi_2}{\pi_2}\right|$$

oraz

$$P = \left|\frac{\pi_1(1-\pi_2) - \pi_2(1-\pi_1)}{\pi_2(1-\pi_1)}\right| = \left|\frac{\pi_1 - \pi_2}{\pi_2(1-\pi_1)}\right|$$

Wiadomo, że:

$$L = \left|\frac{\pi_1 - \pi_2}{\pi_2}\right| \leq \left|\frac{\pi_1 - \pi_2}{\pi_2(1-\pi_1)}\right| = P$$

ponieważ w mianownik prawdodpodobieństwo π_2 jest przemnożone przez wyrażenie $1 - \pi_1$, które jest mniejsze od 1. Zatem mianownik $\pi_2(1-\pi_1) < \pi_2$, co powoduje, że prawa strona nierówności jest ostro większa. W przypadku, gdy $\pi_1, \pi_2 = \frac{1}{2}$, lewa strona nierówności równa się prawej, co należało udowodnić.
Dla dowolnych prawdopodobieństw π_1 oraz π_2 odpowiadających ryzyku w dwóch grupach, wartość ryzyka względnego (RR) jest zawsze bliższa wartości 1 (czyli niezależności) niż odpowiadający jej iloraz szans (OR). Intuicyjnie wynika to z faktu, że OR „przesadza" efekt relacji, szczególnie gdy prawdopodobieństwa są duże — i dlatego jest bardziej oddalony od 1. W analizie danych epidemiologicznych i klinicznych często wskazuje się, że RR jest łatwiejszy do interpretacji, a OR bywa bardziej „drastyczny".

Zadanie 3

Niech D oznacza posiadanie pewnej choroby, a E pozostawanie wystawionym na pewny czynnik ryzyka. W badaniach epidemiologicznych definuje się miarę AR nazywaną ryzykiem przypisanym (ang. attributable risk).
a) Niech $P(E') = 1 - P(E)$, wówczas $AR = [P(D) - P(D|E')]/P(D)$.

- D: posiadanie choroby,
- E: ekspozycja na czynnik ryzyka,
- E': brak ekspozycji,
- $P(D)$: ogólne prawdopodobieństwo zachorowania,

- $P(D \mid E')$: prawdopodobieństwo zachorowania bez czynnika ryzyka.

Miara AR mówi nam, jaki ułamek wszystkich przypadków choroby (D) można przypisać działaniu czynnika ryzyka (E). Licznik - różnica między ogólnym ryzykiem choroby a ryzykiem u osób nieeksponowanych, czyli efekt „ponad tło". Mianownik - skaluje to względem całkowitego ryzyka.

b) Pokaż, żee AR ma związek z ryzykiem względnym, tzn.:

$$AR = [P(E)(RR-1)]/[1+P(E)(RR-1)]$$

$$\text{RR} = \frac{P(D|E)}{P(D|E')} \Rightarrow P(D|E) = RR \cdot P(D|E')$$

$$\begin{aligned} P(D) &= P(E) \cdot P(D|E) + P(E') \cdot P(D|E') \\ &= P(E) \cdot RR \cdot P(D|E') + (1-P(E)) \cdot P(D|E') \\ &= P(D|E') \cdot [P(E) \cdot RR + (1-P(E))] \\ &= P(D|E') \cdot [1+P(E)(RR-1)] \end{aligned}$$

$$\begin{aligned} \text{Licznik AR} &= P(D) - P(D|E') \\ &= P(D|E') \cdot [1+P(E)(RR-1)] - P(D|E') \\ &= P(D|E') \cdot P(E)(RR-1) \end{aligned}$$

$$\begin{aligned} AR &= \frac{P(D)-P(D|E')}{P(D)} \\ &= \frac{P(D|E') \cdot P(E)(RR-1)}{P(D|E') \cdot [1+P(E)(RR-1)]} \\ &= \frac{P(E)(RR-1)}{1+P(E)(RR-1)} \end{aligned}$$

Ryzyko przypisane (AR) określa, jaka część przypadków choroby może być przypisana działaniu badanego czynnika ryzyka. Jest ono funkcją ryzyka względnego (RR) oraz częstości występowania ekspozycji (P(E)). Wzór pokazuje, że nawet jeśli RR jest wysokie, niskie P(E) ogranicza wielkość AR — a więc wpływ czynnika ryzyka na populację.