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 III

Zadanie 7

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

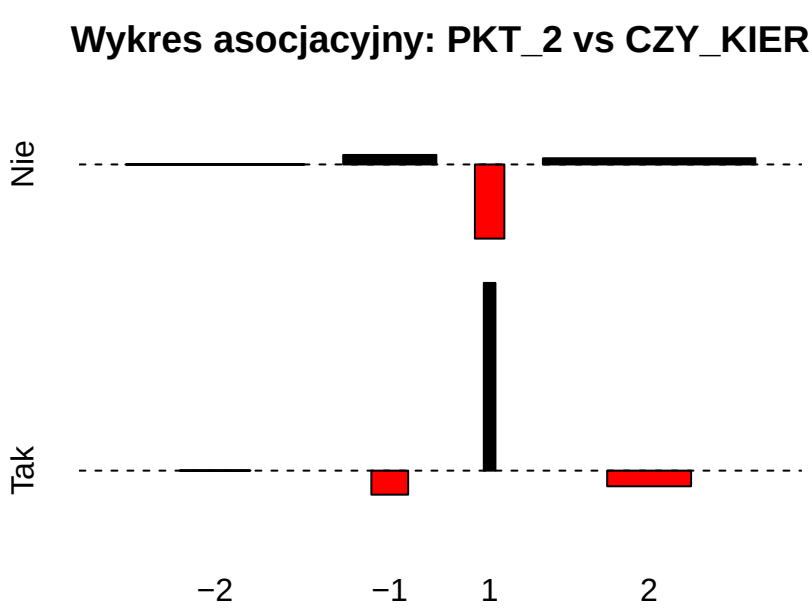
Biorąc poziom istotności $\alpha = 0.05$, nie ma podstaw do odrzucenia hipotezy zerowej o niezależności zmiennych. Oznacza to, że brak jest statystycznie istotnych dowodów na istnienie zależności pomiędzy badanymi cechami.

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).

powrownanie z zad 6!!!

Zadanie 9

Przeprowadzono symulacje w celu oszacowania mocy testu Fishera oraz mocy testu chi--kwadrat Pearsona, generując dane z tabeli 2×2, w której $p_{11} = 1/40$, $p_{12} = 3/40$, $p_{21} = 19/40$, $p_{22} = 17/40$. Symulacje wykonano dla $n = 50$, $n = 100$ oraz $n = 1000$.\

Tabela 5: Porównanie mocy testów chi-kwadrat i Fishera w zależności od liczności próby ($N = 500$, $\alpha = 0{,}05$)

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 (moc 1). 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.

```
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 oraz przyjętym poziomie istotności.

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 $Y_n \sim X_n^2 + N(0,0.1)$. 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.