Analiza danych ankietowych

Sprawozdanie 2

Joanna Kusy Tomasz Srebniak

Spis treści

1 Część I

1.1 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.

Na podstawie odpowiedzi wyznaczono realizację przedziału ufności dla wektora prawdopodobieństw opisującego stopień zadowolenia ze szkolenia. Przyjęto poziom ufności 0.95 oraz zastosowano poprawkę Bonferroniego.

Realizacja przedziału ufności dla wektora prawdopodobieństw wyznaczonego metodą Cloppera–Pearsona.

```
ci_exact <- binom.confint(res, 200, 1 - alpha/5, method='exact')
ci_exact$len <- ci_exact$upper - ci_exact$lower
```

```
##   method   x   n  mean      lower     upper       len
## 1  exact  14 200 0.070 0.03169652 0.1298937 0.0981972
## 2  exact  17 200 0.085 0.04208141 0.1486579 0.1065765
## 3  exact  40 200 0.200 0.13257329 0.2821753 0.1496020
## 4  exact 100 200 0.500 0.40735190 0.5926481 0.1852962
## 5  exact  29 200 0.145 0.08749866 0.2200467 0.1325481
```

Realizacja przedziału ufności dla wektora prawdopodobieństw wyznaczonego metodą Wilsona.

```
##   method   x   n  mean      lower     upper        len
## 1 wilson  14 200 0.070 0.03604773 0.1315662 0.09551852
## 2 wilson  17 200 0.085 0.04660626 0.1500444 0.10343817
## 3 wilson  40 200 0.200 0.13731215 0.2819534 0.14464127
## 4 wilson 100 200 0.500 0.41040470 0.5895953 0.17919061
## 5 wilson  29 200 0.145 0.09228421 0.2205134 0.12822917
```

Realizacja przedziału ufności dla wektora prawdopodobieństw wyznaczonego metodą asymptotyczną, opartą na Centralnym Twierdzeniu Granicznym.

```
##       method   x   n  mean      lower     upper        len
## 1 asymptotic  14 200 0.070 0.02352787 0.1164721 0.09294426
## 2 asymptotic  17 200 0.085 0.03420487 0.1357951 0.10159026
## 3 asymptotic  40 200 0.200 0.12714455 0.2728555 0.14571091
## 4 asymptotic 100 200 0.500 0.40893068 0.5910693 0.18213864
## 5 asymptotic  29 200 0.145 0.08086883 0.2091312 0.12826233
```

Dla zadanych kategorii otrzymujemy proporcje:

- 0.070 – odsetek osób bardzo niezadowolonych,
- 0.085 – odsetek osób niezadowolonych,
- 0.200 – odsetek osób nie mających zdania,

- 0.500 – odsetek osób zadowolonych,
- 0.145 – odsetek osób bardzo zadowolonych.

W zależności od wybranej metody, zmieniają się dolna i górna granica przedziału ufności. Dla kategorii z niewielką liczbą odpowiedzi (14 i 17 głosów) najwęższą realizację przedziału uzyskano przy zastosowaniu metody asymptotycznej. Wraz ze wzrostem liczby odpowiedzi (29 głosów), różnice w długości realizacji przedziałów wyznaczonych metodą asymptotyczną i metodą Wilsona stają się niewielkie. Dla kategorii, w których liczba odpowiedzi wynosiła 40 lub więcej, najwęższa realizacja przedziału została dzięki metodzie Wilsona. Metoda Cloppera–Pearsona nie dała najkrótszej realizacji przedziału dla żadnej z proporcji. Wynika to z faktu, że jest to metoda dokładna, która w sposób konserwatywny utrzymuje zadany poziom ufności.

1.2 Zadanie 2

Napisano funkcję, która wyznacza wartość poziomu krytycznego w testach χ^2 Pearsona oraz χ^2 największej wiarogodności. Testy te służą do weryfikacji hipotezy $H_0 : \mathbf{p} = \mathbf{p_0}$ przy hipotezie alternatywnej $H_1 : \mathbf{p} \neq \mathbf{p_0}$ na podstawie obserwacji $\mathbf{x}$ wektora losowego $\mathbf{X}$ z rozkładu wielomianowego z parametrami n i $\mathbf{p}$.

1.2.1 Test χ^2 Pearsona

Statystyką testową w teście jest

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_{0i})^2}{np_{0i}},$$

gdzie p_{0i} jest i–tą składową wektora $\mathbf{p}$.

Przy założeniu, że H_0 jest prawdziwa statystyka χ^2 ma asymptotycznie rozkład χ^2 z $k-1$ stopniami swobody.

Wartość poziomu krytycznego liczymy jako

$$p\text{–}value = 1 - F_{\chi^2_{k-1}}(\chi^2(\mathbf{x})),$$

gdzie $F_{\chi^2_{k-1}}$ jest dystrybuantą rozkładu χ^2 z $k-1$ stopniami swobody, a $\chi^2(\mathbf{x})$ wartością statystki dla realizacji $\mathbf{x}$.

1.2.2 Test χ^2 największej wiarogodności

Statystyką testową w teście IW jest

$$G^2 = 2\sum_{i=1}^{k} X_i \ln\left(\frac{X_i}{np_{0i}}\right)$$

Przy założeniu, że H_0 jest prawdziwa statystyka G^2 ma asymptotycznie rozkład χ^2 z $k-1$ stopniami swobody.

Wartość poziomu krytycznego liczymy jako

$$p\text{–}value = 1 - F_{\chi^2_{k-1}}(G^2(\mathbf{x})),$$

gdzie $F_{\chi^2_{k-1}}$ jest dystrybuantą rozkładu χ^2 z $k-1$ stopniami swobody, a $G^2(\mathbf{x})$ wartością statystki dla realizacji $\mathbf{x}$.

```
chi_sq_test <- function(x, p0, method='pearson') {
  if (method == 'pearson') {
    n <- sum(x)
    Chi2 <- sum((x - n * p0)^2 / (n * p0))
    p_val <- 1 - pchisq(Chi2, length(x) - 1)
    print(paste("p-value:", p_val))
  }
  if (method == 'IW') {
    n <- sum(x)
    G2 <- 2 * sum(x * log(x / (n * p0)))
    p_val <- 1 - pchisq(G2, length(x) - 1)
    print(paste("p-value:", p_val))
  }
}
```

Dla danych z poprzedniego zadania i $H_0 : \mathbf{p} = (0.1, 0.2, 0.4, 0.2, 0.1)$ za pomocą funkcji otrzymamy następujące wartości poziomu krytycznego:

- dla testu χ^2 Pearsona

```
chi_sq_test(res, c(0.1, 0.2, 0.4, 0.2, 0.1), method='pearson')
```

```
## [1] "p-value: 0"
```

- dla testu χ^2 największej wiarogodności

```
chi_sq_test(res, c(0.1, 0.2, 0.4, 0.2, 0.1), method='IW')
```

```
## [1] "p-value: 0"
```

Dla obu testów dla zadanego poziomu istotności $\alpha = 0.05$ odrzucamy hipotezę zerową H_0 na rzecz hipotezy alternatywnej $H_1 : \mathbf{p} \neq (0.1, 0.2, 0.4, 0.2, 0.1)$.

1.3 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. Przyjęto poziom istotności 0.05. Skorzystano z funkcji napisanej w zadaniu 2.

```
chi_sq_test(vec, rep(1/5, 5), method='pearson')
```

```
## [1] "p-value: 2.75779399316889e-13"
```

```
chi_sq_test(vec, rep(1/5, 5), method='IW')
```

```
## [1] "p-value: 1.07019948458742e-10"
```

Dla obu testów dla zadanego poziomu istotności $\alpha = 0.05$ odrzucamy hipotezę o równomiernym rozkładzie odpowiedzi.

2 Część II

2.1 Zadanie 4

W pakiecie R do wykonania zarówno testu Fishera, jak i testu Freemana – Haltona, można posłużyć się funkcją *fisher.test* z biblioteki *stats*.

Przyjmuje ona następujące argumenty:

- x – dwuwymiarowa tabela kontyngencji w postaci macierzy lub obiektu typu factor,
- y = *NULL* – opcjonalny drugi obiekt typu factor. Używany tylko wtedy, gdy x nie jest macierzą,
- *workspace = 200000* – ilość pamięci roboczej w bajtach przeznaczonej na obliczenia (stosowane tylko dla tabel większych niż 2×2 przy podejściu dokładnym, bez symulacji p-value),
- *hybrid = FALSE* – wartość logiczna określająca, czy zastosować algorytm hybrydowy (dotyczy tylko tabel większych niż 2×2),
- *hybridPars = c(expect = 5, percent = 80, Emin = 1)* – wektor określający “warunki Cochrana” dla poprawności aproksymacji chi–kwadrat,
- *control = list()* – lista z nazwanymi komponentami do niskopoziomowej kontroli algorytmu,
- *or = 1* – wartość hipotetycznego ilorazu szans (stosowane wyłącznie dla tabel 2×2),
- *alternative = “two.sided”* – typ hipotezy alternatywnej (“two.sided”, “less”, “greater”) – stosowane wyłącznie dla tabel 2×2,
- *conf.int = TRUE* – wartość logiczna wskazująca, czy obliczyć przedział ufności dla ilorazu szans (dotyczy tylko tabel 2×2),
- *conf.level = 0.95* – poziom ufności dla przedziału ufności ilorazu szans. Używana tylko przy tabelach o rozmiarach 2x2,
- *simulate.p.value = FALSE* – wartość logiczna określająca, czy *p-value* ma zostać oszacowane za pomocą symulacji Monte Carlo (dotyczy tylko tabel większych niż 2×2),
- *B = 2000* – liczba symulacji Monte Carlo.

Funkcja zwraca obiekt klasy *htest*, który zawiera następujące elementy:

- *p.value* – p wartość testu,
- *conf.int* – przedział ufności dla ilorazu szans (jeśli *conf.int* = *TRUE* i tabela ma rozmiar 2x2),
- *estimate* – estymowana wartość ilorazu szans (tylko dla tabel 2x2),
- *null–value* – wartość ilorazu szans podana w argumencie *or* (tylko dla tabel 2x2),
- *alternative* – typ hipotezy alternatywnej,
- *method* – napis "Fisher's Exact Test for Count Data",
- *data.name* – nazwa danych przekazanych do testu.

2.2 Zadanie 5

Korzystając z testu Fishera, na poziomie istotności 0.05, zweryfikowano hipotezę o niezależności zmiennych **PŁEĆ** i **CZY_KIER**.

```
contingency_table <- table(ankieta$PŁEĆ, ankieta$CZY_KIER)
fisher.test(contingency_table, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  contingency_table
## 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
```

P–value jest większe niż zadany poziom istotności, zatem nie mamy podstaw do odrzucenia hipotezy zerowej. Niezależność jest równoważne równości prawdopodobieństw warunkowych. 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.

2.3 Zadanie 6

Korzystajac z testu Freemana-Haltona na poziomie istotności 0.05 zweryfikowano hipotezy dotyczące omawianych zmiennych.

2.3.1 Podpunkt a

Zajmowanie stanowiska kierowniczego nie zależy od wieku.

```
t <- table(ankieta$CZY_KIER, ankieta$WIEK_KAT)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.7823
## alternative hypothesis: two.sided
```

Na poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Zajmowanie stanowiska kierowniczego i wiek są zmiennymi niezależnymi.

2.3.2 Podpunkt b

Zajmowanie stanowiska kierowniczego nie zależy od stażu pracy.

```
t <- table(ankieta$CZY_KIER, ankieta$STAŻ)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 6.538e-05
## alternative hypothesis: two.sided
```

Na poziomie istotności 0.05 odrzucucamy hipotezę zerową. Zmienne dotyczące zajmowania stanowiska kierowniczego oraz stażu pracy są zależne.

2.3.3 Podpunkt c

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

Zmienne **PYT_2** oraz **CZY_KIER**.

```
t <- table(ankieta$PYT_2, ankieta$CZY_KIER)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.0443
## alternative hypothesis: two.sided
```

Na poziomie istotności 0.05 odrzucamy hipotezę zerową. Zajmowanie stanowiska kierowniczego i stopień zadowolenia ze szkoleń są zmiennymi zależnymi.

Zmienne **CZY_ZADOW** oraz **CZY_KIER**.

```r
t <- table(ankieta$CZY_ZADOW, ankieta$CZY_KIER)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## 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 poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Zajmowanie stanowiska kierowniczego i stopień zadowolenia ze szkoleń są zmiennymi niezależnymi.

2.3.4 Podpunkt d

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

Zmienne **PYT_2** oraz **STAŻ**.

```r
t <- table(ankieta$PYT_2, ankieta$STAŻ)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.01069
## alternative hypothesis: two.sided
```

Na poziomie istotności 0.05 odrzucamy hipotezę zerową. Staż pracy i stopień zadowolenia ze szkoleń są zmiennymi zależnymi.

Zmienne **CZY_ZADOW** oraz **STAŻ**.

```r
t <- table(ankieta$CZY_ZADOW, ankieta$STAŻ)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.4097
```

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

Na poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Staż pracy i stopień zadowolenia ze szkoleń są zmiennymi niezależnymi.

2.3.5 Podpunkt e

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

Zmienne **PYT_2** oraz **PŁEĆ**.

```
t <- table(ankieta$PYT_2, ankieta$PŁEĆ)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.4758
## alternative hypothesis: two.sided
```

Na poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Płeć pracownika i stopień zadowolenia ze szkoleń są zmiennymi niezależnymi.

Zmienne **CZY_ZADOW** oraz **PŁEĆ**.

```
t <- table(ankieta$CZY_ZADOW, ankieta$PŁEĆ)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## 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 poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Płeć pracownika i stopień zadowolenia ze szkoleń są zmiennymi niezależnymi.

2.3.6 Podpunkt f

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

Zmienne **PYT_2** oraz **WIEK_KAT**.

```
t <- table(ankieta$PYT_2, ankieta$WIEK_KAT)
fisher.test(t, conf.level = 0.95, workspace = 2e7)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.3194
## alternative hypothesis: two.sided
```

Na poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Wiek pracownika i stopień zadowolenia ze szkoleń są zmiennymi niezależnymi.

Zmienne **CZY_ZADOW** oraz **WIEK_KAT**.

```
t <- table(ankieta$CZY_ZADOW, ankieta$WIEK_KAT)
fisher.test(t, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t
## p-value = 0.3275
## alternative hypothesis: two.sided
```

Na poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy zerowej. Wiek pracownika i stopień zadowolenia ze szkoleń są zmiennymi niezależnymi.

2.3.7 Porównanie wyników

W zależności od tego, czy do testowania hipotezy wykorzystaliśmy zmienną łączącą kategorie, czy też nie, w podpunktach c i d uzyskaliśmy różne wyniki. Natomiast w podpunktach e i f w obu przypadkach podjęliśmy decyzję o nieodrzuceniu hipotezy zerowej. We wszystkich omawianych przykładach p–value było większe przy wykorzystaniu zmiennej **CZY_ZADOW**.

3 Część III

3.1 Zadanie 7

Do wykonania testu niezależności Chi-kwadrat w R można użyć funkcji *chisq.test* z pakietu *stats*.

Funkcja ta przyjmuje następujące argumenty:

- x – wektor lub macierz,
- y – wektor, ignorowany, jeśli x jest macierzą,

- *correct* – wartość logiczna określająca, czy zastosowana zostanie poprawka Yatesa,
- *p* – wektor prawdopodobieństw,
- *rescale.p* – wartość logiczna określająca, czy przeskalować wektor prawdopodobieństw tak, aby sumował się do 1,
- *simulate.p.value* – wartość logiczna określająca, czy *p-value* oszacowane zostanie za pomocą symulacji Monte Carlo,
- *B* – liczba symulacji Monte Carlo.

Funkcja zwraca obiekt klasy *htest*, który zawiera następujące elementy:

- *statistic* – wartość statystyki testowej,
- *parameter* – liczba stopni swobody,
- *p.value* – p wartość testu,
- *method* – napis mówiący o typie wykonanego testu, określający, czy zastosowano poprawkę Yatesa oraz, czy użyto symulacji Monte Carlo,
- *data.name* – nazwa danych przekazanych do testu,
- *observed* – obserwowane wartości,
- *expected* – oczekiwane wartości przy założeniu hipotezy zerowej,
- *residuals* – reszty Pearsona,
- *stdres* – standaryzowane residua.

3.2 Zadanie 8

Korzystając z funkcji *chisq.test* 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.

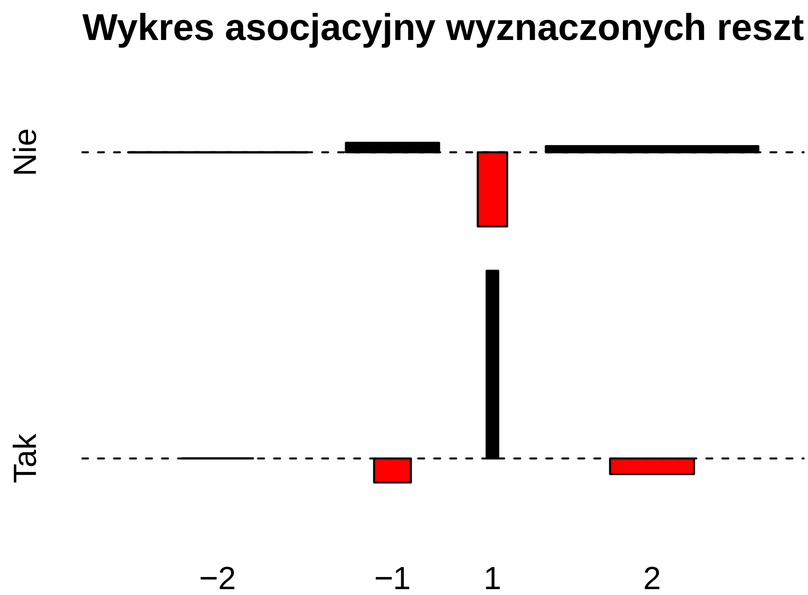
```
t <- table(ankieta$PYT_2, ankieta$CZY_KIER)
chisq.test(t, correct = TRUE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 13.114, df = 3, p-value = 0.004397
```

P–value jest mniejsze niż zadany poziom istotności, zatem odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej. Zajmowane stanowiska kierowniczego i stopień zadowolenia ze szkoleń są zmiennymi zależnymi. Wynik jest taki sam, jak w zadaniu 6c, gdy pod uwagę braliśmy te same zmienne.

Poniżej zaprezentowano wykres asocjacyjny reszt wyznaczonych w teście.

```
assocplot(t, main = "Wykres asocjacyjny wyznaczonych reszt")
```



Z wykresu wynika, że odpowiedzi -1 oraz 2 było więcej wśród osób, które nie zajmują stanowiska kierowniczego. Odpowiedź 1 były bardziej powszechna wśród osób, które zajmują stanowisko kierownicze. Odpowiedź -2 nie była częsciej wskazywana przez żadną z grup.

3.3 Zadanie 9

W pakiecie R do generowania wektorów losowych z rozkładu wielomianowego można posłużyć się funkcją *rmultinom* z pakietu *stats*.

Funkcja ta przyjmuje następujące argumenty:

- n – liczba wektorów losowych,
- *size* – liczba prób,
- *prob* – wektor prawdopodobieństw.

Dla danych wygenerowanych z tabeli, w której $p_{11} = \frac{1}{40}$, $p_{12} = \frac{3}{40}$, $p_{21} = \frac{19}{40}$, $p_{22} = \frac{17}{40}$, przeprowadzano symulacje w celu oszacowania mocy testu Fishera oraz mocy testu chi–kwadrat Pearsona. Symulacje wykonano dla $n \in \{50, 100, 1000\}$.

```
set.seed(123)  # dla powtarzalności

# Parametry
prob <- c(1, 3, 19, 17) / 40
alpha <- 0.05
n_iter <- 10000
n_values <- c(50, 100, 1000)
```

Funkcja *simulate_power* przeprowadza symulacje dla zadanych parametrów oblicza moce rozważanych testów. W każdej iteracji generowany jest wektor losowy z rozkładu wielomianowego, na podstawie którego, za pomocą funkcji bibliotecznych, obliczane są p–values. Dzieląc liczbę odrzuconych hipotez zerowych przez liczbę iteracji, otrzymujemy estymowaną wartość mocy testu.

```r
# Funkcja do przeprowadzenia symulacji
simulate_power <- function(n, iter, prob, alpha) {
  fisher_rejections <- 0
  chisq_rejections <- 0

  for (i in 1:iter) {
    sample <- rmultinom(1, size = n, prob = prob)
    table <- matrix(sample, nrow = 2, byrow = TRUE)

    fisher_p <- fisher.test(table)$p.value
    chisq_p <- suppressWarnings(
      chisq.test(table, correct = FALSE, simulate.p.value = TRUE)$p.value
      )

    iter_chisq <- iter

    if (fisher_p < alpha) {
      fisher_rejections <- fisher_rejections + 1
    }
    if (!is.na(chisq_p) && chisq_p < alpha) {
      chisq_rejections <- chisq_rejections + 1
    }
    if (is.na(chisq_p)) {
      iter_chisq <- iter_chisq - 1
    }
  }
  fisher_power <- fisher_rejections / iter
  chisq_power <- chisq_rejections / iter_chisq

  return(c(Fisher = fisher_power, ChiSq = chisq_power))
}
```

Poniżej przedstawiono wyniki symulacji.

```
##        n = 50 n = 100 n = 1000
## Fisher 0.1174  0.3161   0.9995
## ChiSq  0.1199  0.3145   0.9995
```

Wraz ze wzrostem wartości n rośnie moc testu. Dla $n = 50$ moc testu chi–kwadrat Pearsona jest większa od mocy testu Fishera. Dla $n = 100$ jest odwrotnie. Przy $n = 1000$ moce testów są równe z dokładnością do 4 miejsc po przecinku.

3.4 Zadanie 10

Napisano funkcję, która dla danych z tablicy dwudzielczej oblicza wartość poziomu krytycznego w teście niezależności opartym na ilorazie wiarogodności.

Statystyką testową w teście jest

$$G^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right),$$

gdzie $O_{i,j}$ to obserowana wartość, a $E_{i,j}$ to wartość spodziewana policzona według wzoru $E_{i,j} = \frac{O_{i+}O_{+j}}{n}$.

P–value obliczymy posługując się wzorem

$$p\text{–}value = 1 - F_{\chi^2_{(R-1)(C-1)}}(G^2),$$

gdzie $F_{\chi^2_{(R-1)(C-1)}}$ jest dystrybuantą rozkładu χ^2 z $(R-1)(C-1)$ stopniami swobody.

Na podstawie opisanego schematu funkcja *lr_test_p–value* oblicza wartość poziomu krytycznego w teście niezależności opartym na ilorazie wiarogodności.

```
lr_test_pvalue <- function(table) {
  # Obserwacje
  observed <- table
  total <- sum(observed)

  # Oczekiwane liczności pod H0 (niezależność)
  row_totals <- rowSums(observed)
  col_totals <- colSums(observed)
  expected <- outer(row_totals, col_totals) / total

  # Wykluczenie zer (dla bezpieczeństwa w log)
  valid <- observed > 0

  # Obliczenie statystyki G^2
  G2 <- 2 * sum(observed[valid] * log(observed[valid] / expected[valid]))

  # Stopnie swobody: (r-1)(c-1)
  df <- (nrow(observed) - 1) * (ncol(observed) - 1)

  # Obliczenie p-value
  p_value <- 1 - pchisq(G2, df = df)
```

```r
  return(p_value)
}
```

Korzystając z napisanej funkcji, wykonano test dla danych przeanalizowanych w zadaniu 8.

```r
t <- table(ankieta$PYT_2, ankieta$CZY_KIER)
lr_test_pvalue(t)
```

```
## [1] 0.03968956
```

Przyjmjując za poziom istotności 0.05, odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej. Zajmowane stanowisko kierownicze i stopień zadowolenia ze szkoleń są zmiennymi zależnymi. Wynik jest taki sam, jak w zadaniu 8.