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
exact	14	200	0.070	0.0316965	0.1298937	0.0981972
exact	17	200	0.085	0.0420814	0.1486579	0.1065765
exact	40	200	0.200	0.1325733	0.2821753	0.1496020
exact	100	200	0.500	0.4073519	0.5926481	0.1852962
exact	29	200	0.145	0.0874987	0.2200467	0.1325481

Realizacja przedziału ufności dla wektora prawdopodobieństw wyznaczonego metodą Wilsona.

method	x	n	mean	lower	upper	len
wilson	14	200	0.070	0.0360477	0.1315662	0.0955185
wilson	17	200	0.085	0.0466063	0.1500444	0.1034382
wilson	40	200	0.200	0.1373121	0.2819534	0.1446413
wilson	100	200	0.500	0.4104047	0.5895953	0.1791906
wilson	29	200	0.145	0.0922842	0.2205134	0.1282292

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

method	x	n	mean	lower	upper	len
asymptotic	14	200	0.070	0.0235279	0.1164721	0.0929443
asymptotic	17	200	0.085	0.0342049	0.1357951	0.1015903
asymptotic	40	200	0.200	0.1271445	0.2728555	0.1457109
asymptotic	100	200	0.500	0.4089307	0.5910693	0.1821386
asymptotic	29	200	0.145	0.0808688	0.2091312	0.1282623

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 **x** wektora losowego **X** z rozkładu wielomianowego z parametrami n i **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 **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 **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. Zakładamy, że 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. Zakładamy, że 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. Zakładamy, że 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. Zakładamy, że 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. Zakładamy, że 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. Zakładamy, że 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. Zakładamy, że 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. Zakładamy, że 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. Zakładamy, że 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. Zakładamy, że 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.

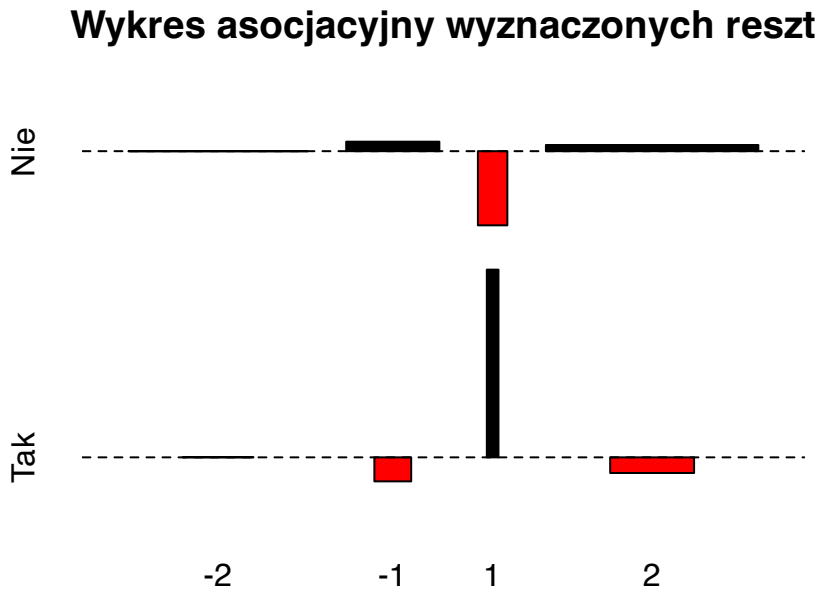
```
t <- table(ankieta$PYT_2, ankieta$CZY_KIER)
chisq.test(t, correct = TRUE)
```

```
##
##  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. Zakładamy, że zajmowanie 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 częściej 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
```

```r
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)

  return(p_value)
}
```

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

```
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. Zakładamy, że zajmowane stanowisko kierownicze i stopień zadowolenia ze szkoleń są zmiennymi zależnymi. Wynik jest taki sam, jak w zadaniu 8.

4 Część IV i V

4.1 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.

	Rak płuc	Choroba niedokrwienna serca
Palenie	0.0014	0.00669
Nie palenie	0.0001	0.00413

4.1.1 Różnica proporcji

Różnica proporcji dla palaczy i osób niepalących wynosi 0.0013 dla raka płuc oraz 0.00256 dla choroby niedokrwiennej serca. Obie wartości są bliskie 0, które jest równoważne z niezależnością zmiennych. Warto jednak zauważyć, że różnica proporcji dla choroby niedokrwiennej serca jest ponad 2 razy większa niż dla raka płuc.

4.1.2 Ryzyko względne

Ryzyko względne dla palaczy i osób niepalących wynosi 14 dla raka płuc oraz około 1.6199 dla choroby niedokrwiennej serca. Obie wartości są większe od 1, co wskazuje na związek zmiennych. Ryzyko zgonu na raka płuc jest ponad 14 razy większe dla palaczy niż dla osób niepalących. Dla choroby niedokrwiennej serca ryzyko zgonu dla palacza jest większe o około 1.62.

4.1.3 Iloraz szans

Iloraz szans dla palaczy i osób niepalących wynosi około 14.0182 dla raka płuc oraz 1.624 dla choroby niedokrwiennej serca. Oba ilorazy są większe od 1, co oznacza, że szansa zgonu z powodu tych chorób jest wyższa wśród palaczy niż wśród osób niepalących.. Większe odchylenie od 1 wskazuje na silniejszy związek, co w tym przypadku ma miejsce dla raka płuc.

4.2 Zadanie 12

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

	śmiertelny	nieśmiertelny
bez pasów	1085	55623
z pasami	703	441239

4.2.1 Podpunkt a

Warunkowe prawdopodobieństwo śmierci w wypadku ze względu na użycie przez kierowców i pasażerów pasa bezpieczeństwa wynosi 0.0015907.

Warunkowe prawdopodobieństwo śmierci w wypadku ze względu na nieużycie przez kierowców i pasażerów pasa bezpieczeństwa wynosi 0.0191331.

4.2.2 Podpunkt b

Warunkowe prawdopodobieństwo użycia pasa bezpieczeństwa ze względu na kierowców i pasażerów ze śmiertelnymi obrażeniami wynosi 0.3931767.

Warunkowe prawdopodobieństwo użycia pasa bezpieczeństwa ze względu na kierowców i pasażerów, którzy przeżyli wypadek wynosi 0.8880514.

4.2.3 Podpunkt c

Naturalnym wyborem zmiennej objaśniającej jest zmienna dotycząca użycia pasa bezpieczeństwa. Poprzez taki wybór traktujemy śmierć w wypadku jako zmienną objaśnianą.

4.2.3.1 Różnica proporcji

Różnica proporcji dla osób z niezapiętymi i zapiętymi pasami wynosi 0.0175424. Wartość ta jest nieweiele większa od 0, co może wskazywać na słabą zależność zmiennych.

4.2.3.2 Ryzyko względne

Ryzyko względne dla osób z niezapiętymi i zapiętymi pasami wynosi 12.0280537. Wartość ta jest większa od 1, co wskazuje na związek zmiennych. Ryzyko śmierci w wypadku dla osób z niezapiętymi pasami jest około 12.0281 razy większe niż dla osób z zapiętymi pasami.

4.2.3.3 Iloraz szans

Iloraz szans dla osób z niezapiętymi i zapiętymi pasami wynosi 12.2432 i jest większy od 1, co oznacza, że szansa śmierci w wypadku jest większa dla osób bez pasów niż dla osób z zapiętym pasem bezpieczeństwa.

Ryzyko względne oraz iloraz szans przyjmują podobne wartości ze względu na małą liczbę śmiertelnych wypadków.

4.3 Zadanie 13

Obliczono wartości odpowiednich miar współzmienności dla niektórych z wcześniej omawianych zmiennych.

4.3.1 Podpunkt a

Stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie (zmienna CZY_ZADOW) i zajmowane stanowisko.

```
t <- table(ankieta$CZY_ZADOW, ankieta$CZY_KIER)
GoodmanKruskalTau(t)
```

```
## [1] 0.0004091802
```

Wartość współczynnika τ jest bliska zeru, co wskazuje na brak istotnej zależności między zmiennymi. Dodatkowo w zadaniu 6c, na podstawie testu Freemana–Haltona przy poziomie istotności 0.05, nie odrzuciliśmy hipotezy o niezależności. Uwzględniając oba wyniki, można wnioskować o braku zależności między analizowanymi zmiennymi.

4.3.2 Podpunkt b

Stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie i staż pracy.

4.3.2.1 Współczynnik τ

Jako zmienną mierzącą stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie przyjęto zmienną **CZY_ZADOW**.

```
t <- table(ankieta$CZY_ZADOW, ankieta$STAŻ)
GoodmanKruskalTau(t)
```

```
## [1] 0.008886788
```

Wartość współczynnika τ jest bliska zeru, co wskazuje na brak istotnej zależności między zmiennymi. Dodatkowo w zadaniu 6d, na podstawie testu Freemana–Haltona przy poziomie istotności 0.05, nie odrzuciliśmy hipotezy o niezależności. Uwzględniając oba wyniki, można wnioskować o braku zależności między analizowanymi zmiennymi.

4.3.2.2 Współczynnik γ

Jako zmienną mierzącą stopień zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie przyjęto zmienną **PYT_2**.

```
t <- table(ankieta$PYT_2, ankieta$STAŻ)
GoodmanKruskalGamma(t)
```

```
## [1] 0.09084262
```

Wartość współczynnika γ jest większa od zera, co wskazuje na dodatnią zależność między zmiennymi. Oznacza to, że wyższe kategorie jednej zmiennej mają tendencję do współwystępowania z wyższymi kategoriami drugiej zmiennej. Dodatkowo, w zadaniu 6d na podstawie testu Freemana–Haltona przy poziomie istotności 0.05, odrzuciliśmy hipotezę o niezależności zmiennych. Uwzględniając oba wyniki, można wnioskować o istnieniu zależności między analizowanymi zmiennymi.

4.3.3 Podpunkt c

Zajmowane stanowisko i staż pracy.

```
t <- table(ankieta$DZIAŁ, ankieta$STAŻ)
GoodmanKruskalTau(t)
```

```
## [1] 0.1226655
```

Wartość współczynnika τ jest większa od zera, co wskazuje na zależności między zmiennymi. Dodatkowo w zadaniu 6b, na podstawie testu Freemana–Haltona przy poziomie istotności 0.05, odrzuciliśmy hipotezę o niezależności zmiennych. Niskie p-value świadczy o statystycznie istotnej zależności między zmiennymi. Uwzględniając oba wyniki, można wnioskować o istnieniu zależności między analizowanymi zmiennymi.

4.4 Zadanie 14

Napisz własną funkcję do przeprowadzania analizy korespondencji. Funkcja jako argument przyjmuje tablicę dwudzielczą i zwraca obliczone wartości odpowiednich wektorów i macierzy, współrzędnych punktów oraz odpowiedni wykres.

```
correspondence_analysis <- function(t, variables = c("row", "col")) {
  # macierz korespondencji
  P <- prop.table(t)
  # wektory częstości brzegowych
  r <- rowSums(P)
  c <- colSums(P)
```

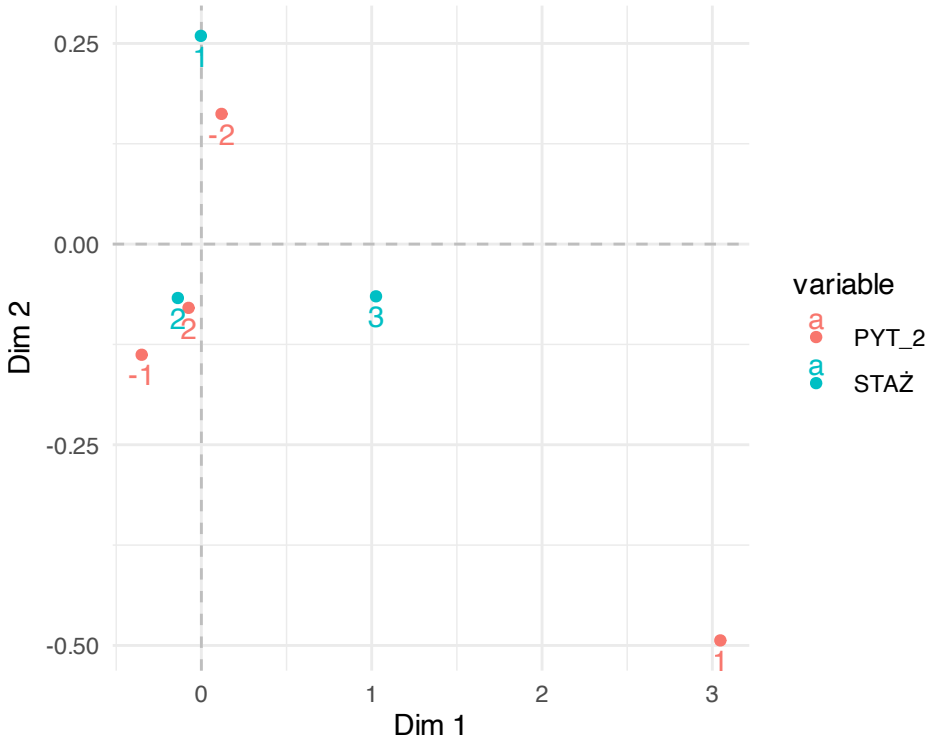
```r
# macierz częstości wierszowych
D_r <- diag(r)
# macierz profili wierszowych
R <- solve(D_r) %*% P
# macierz częstości kolumnowych
D_c <- diag(c)
# macierz profili kolumnowych
C <- P %*% solve(D_c)

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

# dekompozycja według wartości osobliwych macierzy
SVD <- svd(A)
# macierze ortogonalne
U <- SVD$u
V <- SVD$v
# macierz diagonalna o niezerowych wartościach osobliwych macierzy A
# ułożonych w porządku nierosnącym
D <- diag(SVD$d)

# współrzędne kategorii cech dla wierszy
F <- solve(sqrtm(D_r)) %*% U %*% D
# współrzędne kategorii cech dla kolumn
G <- solve(sqrtm(D_c)) %*% V %*% D
print(F)
print(G)

# rysowanie wykresu, typ: principal
row_std <- F[, 1:2]
col_std <- G[, 1:2]

ggplot(data.frame(x = c(row_std[, 1], col_std[, 1]),
                  y = c(row_std[, 2], col_std[, 2]),
                  label = c(rownames(t), colnames(t)),
                  variable = c(rep(variables[1], nrow(t)),
                               rep(variables[2], ncol(t)))),
       aes(x = x, y = y, color = variable)) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "grey") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "grey") +
  geom_point(aes(x = x, y = y)) +
  geom_text(aes(x = x, y = y, label = label), vjust = 1.5) +
  labs(x = "Dim 1", y = "Dim 2") +
  theme_minimal()
```

```
}
```

Korzystając z napisanej funkcji wykonano analizę korespondencji dla danych dotyczących stopnia zadowolenia ze szkoleń w kontekście dopasowania do indywidualnych potrzeb w pierwszym badanym okresie oraz stażu pracy.

```
t <- table(ankieta$PYT_2, ankieta$STAŻ)
correspondence_analysis(t, variables = c("PYT_2", "STAŻ"))
```

```
##              [,1]       [,2]          [,3]
## [1,]  0.11826617  0.1622275 -2.593960e-17
## [2,] -0.35056182 -0.1378599 -6.255756e-17
## [3,]  3.04672664 -0.4937333 -4.883933e-17
## [4,] -0.07532609 -0.0794247 -1.097916e-17
##               [,1]        [,2]          [,3]
## [1,] -0.003060932  0.25943499 -2.695995e-17
## [2,] -0.138246087 -0.06714975 -2.695995e-17
## [3,]  1.025260549 -0.06504575 -2.695995e-17
```



4.4.1 Profile wierszy

Z wykresu wynika, że profile dla odpowiedzi -1 oraz 2 są do siebie zbliżone. Obie znajdują się blisko początku układu współrzędnych, co sugeruje, że ich rozkład warunkowy jest zbliżony do rozkładu oczekiwanego przy niezależności zmiennych. Odpowiedź -2 również znajduje się blisko środka, ale nie tworzy wyraźnego skupiska z innymi punktami. Najbardziej wyróżnia się odpowiedź 1, która leży najdalej od środka i pozostałych kategorii, co wskazuje na jej odmienny profil.

4.4.2 Profile kolumn

Z wykresu wynika, że profil odpowiedzi dla liczby lat stażu oznaczonego cyfrą 2 jest tym, który najbardziej przypomina rozkład oczekiwany przy niezależności zmiennych. W przypadku pozostałych kategorii punkty są bardziej oddalone od środka układu współrzędnych, co wskazuje na większe różnice między obserwowanymi a oczekiwanymi rozkładami. Żadna para punktów nie jest blisko siebie, co sugeruje, że ich rozkład warunkowy nie jest podobny.

5 Zadania dodatkowe

5.1 Zadanie *1

Napisano funkcję, która dla dwóch wektorów danych oblicza wartość poziomu krytycznego (p--value) w teście opartym na korelacji odległości.

Statystyką testową w teście jest wartość

$$dCor = \frac{\sqrt{\sum_{i=1}^{R}\sum_{j=1}^{C}\left(p_{ij}-p_{i+}p_{+j}\right)^2}}{\left(\sum_{i=1}^{R}p_{i+}^2\left(\sum_{i=1}^{R}p_{i+}^2+1\right)-2\sum_{i=1}^{R}p_{i+}^3\right)^{1/4}\cdot\left(\sum_{j=1}^{C}p_{+j}^2\left(\sum_{j=1}^{C}p_{+j}^2+1\right)-2\sum_{j=1}^{C}p_{+j}^3\right)^{1/4}}$$

gdzie: - p_{ij} to prawdopodobieństwo wystąpienia zdarzenia i i j, - p_{i+} to rozkład brzegowy zmiennej x, - p_{+j} to rozkład brzegowy zmiennej y, - R to liczba kategorii zmiennej x, - C to liczba kategorii zmiennej y.

Funkcja *dCor_pvalue* oblicza statystykę testową, a następnie wykonuje 1000 permutacji. P–value jest obliczane jest jako liczba permutacji, dla których wartość statystyki testowej jest większa lub równa wartości statystyki testowej podzielona przez liczbę wszystkich permutacji.

```
dCor_pvalue <- function(x, y) {

  dCor <- function(x, y) {
    observed <- table(x, y)
    total <- sum(observed)

    probs <- prop.table(observed)
    row_probs <- rowSums(probs)
    col_probs <- colSums(probs)

    numerator <- 0

    for (i in 1:nrow(observed)) {
      for (j in 1:ncol(observed)) {
        numerator <- numerator + (probs[i, j]
```

```
                                  - row_probs[i] * col_probs[j])^2
      }
    }
    numerator <- sqrt(numerator)

    denominator <- (
      sum(row_probs^2 * (sum(row_probs^2) + 1)) - 2*sum(row_probs^3)
      )^0.25*
      (
        sum(col_probs^2 * (sum(col_probs^2) + 1)) - 2*sum(col_probs^3)
        )^0.25
    return(numerator / denominator)
  }

  statistic <- dCor(x, y)

  for (i in 1:1000) {
    x_perm <- sample(x)
    y_perm <- sample(y)

    statistic_perm <- dCor(x_perm, y_perm)

    if (statistic_perm >= statistic) {
      p_value <- (i + 1) / 1000
      break
    }
  }
  return(p_value)
}
```

Dla zmiennych **PYT_2** i **STAŻ** zweryfikowano hipotezę o niezależności przy użyciu napisanej funkcji.

```
dCor_pvalue(ankieta$PYT_2, ankieta$CZY_KIER)
```

```
## [1] 0.002
```

Na poziomie istotności 0.05 odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej. Zakładamy, że zajmowane stanowisko kierownicze i stopień zadowolenia ze szkoleń są zmiennymi zależnymi.

5.2 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).

Wzór na ryzyko względne to $\frac{\pi_1}{\pi_2}$, a wzór na iloraz szans to $\frac{\pi_1}{1-\pi_1} / \frac{\pi_2}{1-\pi_2}$.

5.2.1 Przypadek 1: $\pi_1 = \pi_2$

W przypadku, gdy $\pi_1 = \pi_2$, zarówno ryzyko względne, jak i iloraz szans, są równe 1. Zatem wartość ryzyka względnego nie jest bardziej oddalona od wartości 1 niż wartość odpowiadającego ilorazu szans.

5.2.2 Przypadek 2: $\pi_1 > \pi_2$

W przypadku, gdy $\pi_1 > \pi_2$, ryzyko względne jest większe od 1. Iloraz szans można rozpisać jako:

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

Zatem jest to ryzyko względne pomnożone przez dodatkowy czynnik. Jako, że $\pi_2 < \pi_1$, to $1-\pi_2 > 1-\pi_1$. Zatem iloraz szans jest większy od ryzyka względnego i tym samym bardziej oddalony od 1.

5.2.3 Przypadek 3: $\pi_1 < \pi_2$

Rozważając $\pi_1 < \pi_2$, ryzyko względne jest mniejsze od 1. Iloraz szans można rozpisać analogicznie, jak w przypadku 2. Tym razem $1-\pi_2 < 1-\pi_1$. Zatem iloraz szans jest mniejszy od ryzyka względnego i tym samym bardziej oddalony od 1.

5.3 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*).

5.3.1 Interpretacja

Niech $P(E') = 1 - P(E)$, wówczas $AR = [P(D) - P(D|E')]/P(D)$.

Na podstawie wzoru możemy stwierdzić, że ryzyko przypisane (*AR*) wyraża część ogólnego ryzyka *P(D)*, którą można przypisać ekspozycji *E*. Wartość *AR* = 0 oznacza, że nie ma związku między czynnikiem ryzyka a chorobą. Wartość *AR* = 1 oznacza, że wszystkie przypadki choroby są spowodowane czynnikiem ryzyka.

5.3.2 Związek z ryzykiem względnym

Pokażemy, że $AR = [P(E)(RR-1)]/[1+P(E)(RR-1)]$.

RR możemy wyrazić wzorem

$$RR = \frac{P(D|E)}{P(D|E')}. \tag{1}$$

A ogólne ryzyko zapisać jako

$$\begin{aligned} P(D) &= P(D \mid E) \cdot P(E) + P(D \mid E') \cdot P(E') \\ &\overset{(1)}{=} RR \cdot P(D \mid E') \cdot P(E) + P(D \mid E') \cdot (1 - P(E)) \\ &= P(D \mid E') \left(P(E)(RR - 1) + 1\right). \end{aligned} \tag{2}$$

Zatem

$$AR = \frac{P(D) - P(D|E')}{P(D)} \overset{(2)}{=} \frac{P(D|E') \left(P(E)(RR-1) + 1 - 1\right)}{P(D|E') \left(P(E)(RR-1) + 1\right)} = \frac{P(E)(RR-1)}{1 + P(E)(RR-1)}.$$