Analiza danych ankietowych

Sprawozdanie 1

Weronika Jaszkiewicz Weronika Pyrtak

Spis treści

Część I

Zadanie 1

W pewnej dużej firmie technologicznej przeprowadzono ankietę mającą na celu ocenę skuteczności programów szkoleniowych dla pracowników. Wzięło w niej udział 200 losowo wybranych osób (losowanie proste ze zwracaniem).

Zadanie 1

Wczytywanie danych z pliku ankieta.csv.

```
df <- read.csv("ankieta.csv", sep = ";", fileEncoding = "Latin2")
colnames(df) <- c('DZIAŁ','STAŻ','CZY_KIER', 'PYT_1',
                  'PYT_2', 'PYT_3', 'PŁEĆ', 'WIEK')

attach(df)
```

Powyższe dane zawierają 200 wierszy oraz 8 kolumn.

Następnie sprawdzono typy przyjmowanych zmiennych.

Tabela 1: Typy zmiennych

DZIAŁ	character
STAŻ	integer
CZY_KIER	character
PYT_1	integer
PYT_2	integer
PYT_3	integer
PŁEĆ	character
WIEK	integer

Zamieniono zmienne o typie *character* na typ *factor*.

Przeszukano zbiór pod względem braków danych.

Liczba wartości brakujących wynosi: 0

Sprawdzono, czy typy zmiennych zostały prawidłowo rozpoznane.

1. zmienne ilościowe (typ numeric)

Tabela 2: Zmienne ilościowe

STAŻ	2
PYT_1	4
PYT_2	5
PYT_3	6
WIEK	8

Liczba zmiennych ilościowych: 5

2. zmienne jakościowe (typ factor)

Tabela 3: Zmienne jakościowe (factor)

DZIAŁ	1
CZY_KIER	3
PŁEĆ	7

Liczba zmiennych jakościowych (typ factor): 3

Zadanie 2

Utworzono zmienna "WIEK_KAT" przeprowadzając kategoryzację zmiennej "WIEK" korzystając z następujących przedziałów do 35 lat, między 36 a 45 lat, między 46 a 55 lat, powyżej 55 lat.

```
df$WIEK_KAT <- cut(df$WIEK,
breaks = c(-Inf, 35, 45, 55, Inf),
labels = c('0-35', '36-45', '46-55', '55+'))
```

Zadanie 3

Sporządzono tablice liczności dla zmiennych: DZIAŁ, STAŻ, CZY_KIER, PŁEĆ, WIEK_KAT.

Tabela 4: Tablica ilości dla zmiennej DZIAŁ

	HR	IT	MK	PD
Ilość	31	26	45	98

Tabela 5: Tablica ilości dla zmiennej STAŻ

	<1 rok	1-2 lata	3+ lat
Ilość	41	140	19

Tabela 6: Tablica ilości dla zmiennej CZY KIER

	TAK	NIE
Ilość	173	27

Tabela 7: Tablica ilości dla zmiennej PŁEĆ

	Kobieta	Mężczyzna
Ilość	71	129

Tabela 8: Tablica ilości dla zmiennej WIEK KAT

	0-35	36-45	46-55	55+
Ilość	26	104	45	25

Największą grupę pracowników stanowi dział PD (98 osób), a najmniejszą IT (26 osób). Działy HR i MK mają pośrednie wartości (odpowiednio 31 i 45 osób). Może to wskazywać na różne zapotrzebowanie na pracowników w poszczególnych działach.

Większość pracowników ma od 1 do 2 lat stażu. 41 osób pracuje krócej niż rok, a tylko 19 osób ma staż powyżej 3 lat. Wskazuje to na dużą rotację pracowników lub na to, że firma stosunkowo niedawno zatrudniła większość obecnej kadry.

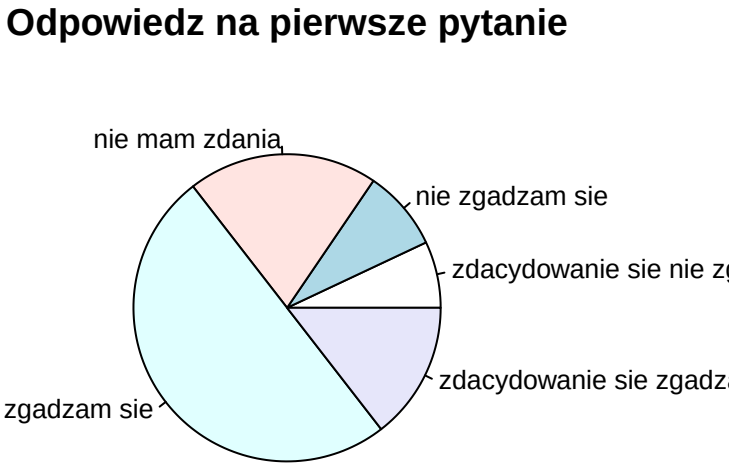
Wielu pracowników (173 osoby) pełni funkcje kierownicze. Może to oznaczać, że w firmie jest dużo takich stanowisk lub że definicja „kierownika" obejmuje szerokie spektrum stanowisk.

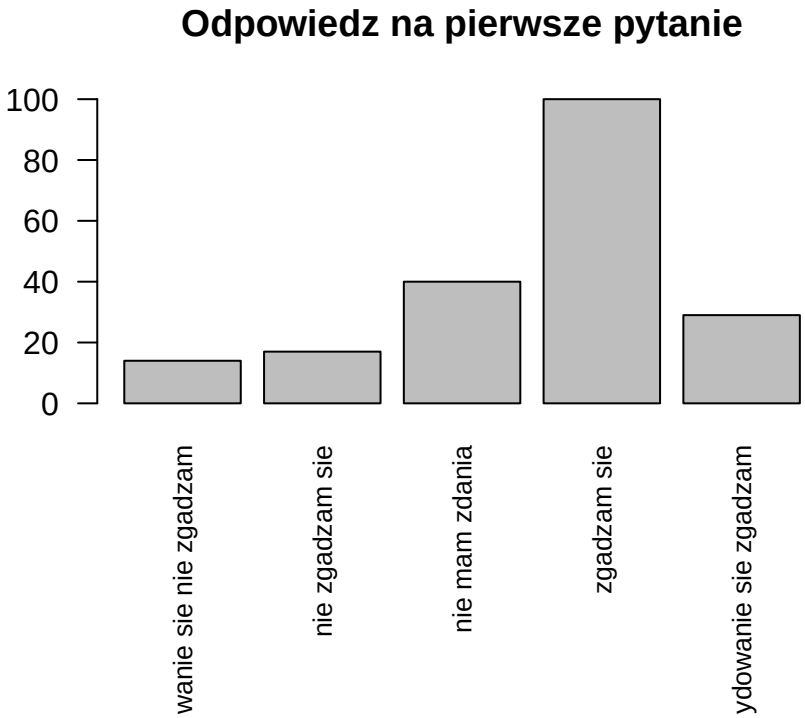
W firmie przeważają mężczyźni (129 osób) nad kobietami (71 osób). Może to wynikać z charakteru działalności firmy lub preferencji rekrutacyjnych. Najwięcej pracowników jest w grupie wiekowej 36-45 lat. Pozostałe grupy wiekowe mają znacznie mniejszą reprezentację.

Jest 5 zmiennych ilościowych (np. staż, wiek, odpowiedzi na pytania PYT_1–PYT_3) i 4 zmienne jakościowe, np. dział, płeć, kategoria wieku. Brak wartości brakujących wskazuje na kompletność i dobrą jakość danych.

Zadanie 4

Sporządzono wykresy kołowe oraz wykresy słupkowe dla zmiennych: PYT_1 (Jak bardzo zgadzasz się ze stwierdzeniem, że firma zapewnia odpowiednie wsparcie i materiały umożliwiające skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń?) oraz PYT_2 (Jak bardzo zgadzasz się ze stwierdzeniem, że firma oferuje szkolenia dostosowane do twoich potrzeb, wspierając twój rozwój zawodowy i szanse na awans?).

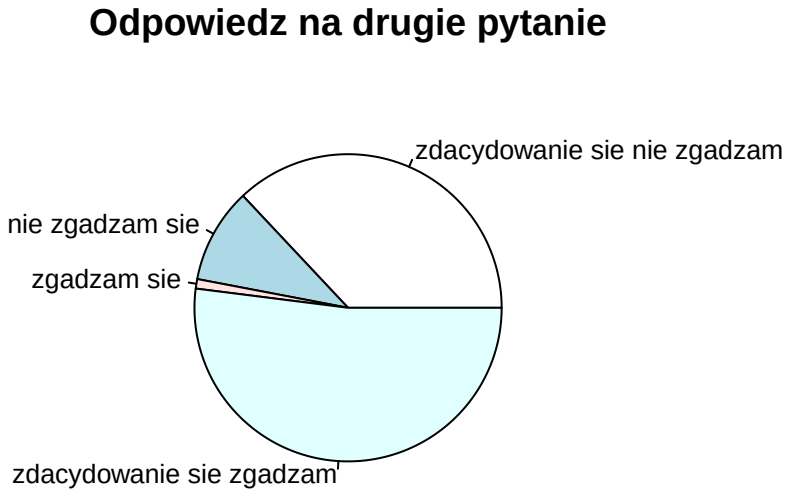


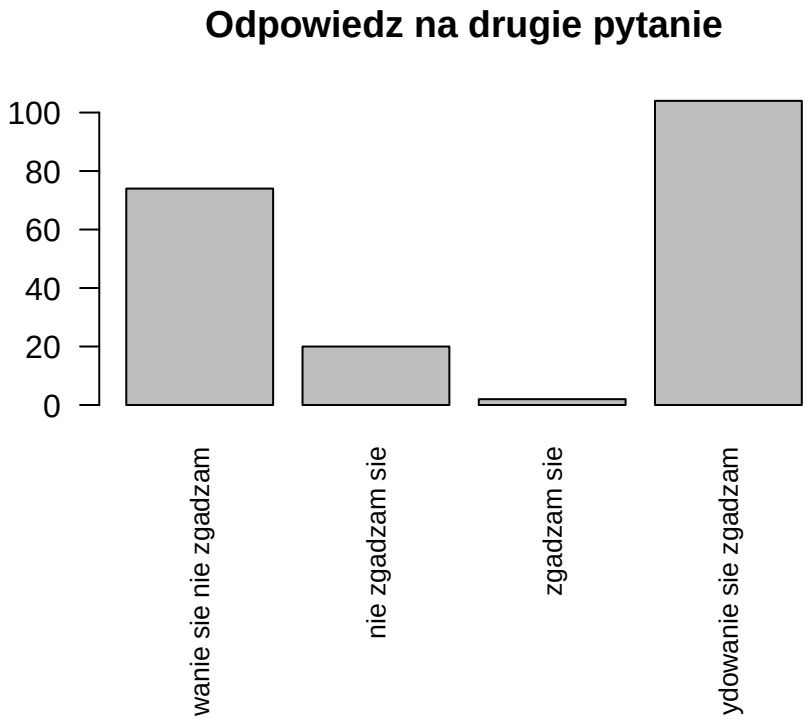


Największa liczba respondentów zaznaczyła odpowiedź "zgadzam się", co wskazuje, że ogólna ocena wsparcia i materiałów dostarczanych przez firmę jest pozytywna. Istnieje także pewna grupa osób, które "zdecydowanie się zgadzają", co podkreśla, że część pracowników uważa wsparcie za bardzo dobre.

Znaczna część pracowników wybrała odpowiedź "nie mam zdania", co może sugerować, że nie mieli oni okazji skorzystać ze wsparcia lub materiały nie są dla nich wystarczająco widoczne.

Pewna część respondentów zaznaczyła opcje "nie zgadzam się" i "zdecydowanie się nie zgadzam", ale ich liczba jest stosunkowo mała w porównaniu do grupy zadowolonych pracowników. Może to wskazywać na pewne niedociągnięcia w dostępie do materiałów lub ich jakości, jednak nie jest to powszechny problem.





Największa część respondentów zaznaczyła odpowiedź "zdecydowanie się zgadzam", co wskazuje, że większość pracowników uważa szkolenia oferowane przez firmę za dobrze dopasowane do ich potrzeb. Istnieje także spora grupa osób, które wybrały opcję "zgadzam się", co dodatkowo potwierdza ogólnie pozytywne nastawienie wobec polityki szkoleniowej firmy.

Stosunkowo niewielka część respondentów wybrała odpowiedź "nie zgadzam się" oraz "zdecydowanie się nie zgadzam", co sugeruje, że niektóre osoby mogą mieć trudności z dostępem do odpowiednich szkoleń lub nie uważają ich za skuteczne w kontekście swojego rozwoju zawodowego.

Pomimo że liczba negatywnych odpowiedzi jest niewielka, warto przeanalizować, czy istnieją konkretne obszary, w których szkolenia mogą być bardziej dostosowane do indywidualnych potrzeb.

Zadanie 5

Sporządzono tablice wielodzielcze dla par zmiennych: PYT_1 i DZIAŁ, PYT_1 i STAŻ, PYT_1 i CZY_KIER, PYT_1 i PŁEĆ oraz PYT_1 i WIEK_KAT.

Tabela 9: Tabela wielodzielcza dla zmiennych PYT 1 i DZIAL

	HR	IT	MK	PD
zdacydowanie się nie zgadzam	2	0	3	9
nie zgadzam się	2	2	3	10
nie mam zdania	5	4	14	17
zgadzam się	19	15	15	51
zdacydowanie się zgadzam	3	5	10	11

Najbardziej pozytywne opinie pochodzą z działów HR i PD – większość respondentów z tych działów wybrała opcje "zgadzam się" i "zdecydowanie się zgadzam". Dział MK (Marketing) jest bardziej podzielony – występuje większy odsetek osób, które nie mają zdania, co może sugerować, że dla tej grupy pytanie nie było jednoznaczne lub temat ich mniej dotyczył. Dział IT wykazuje najmniejszy poziom negatywnych odpowiedzi ("zdecydowanie się nie zgadzam" = 0), ale też stosunkowo niewiele osób udzieliło odpowiedzi skrajnie pozytywnej.

Tabela 10: Tabela wielodzielcza dla zmiennych PYT 1 i STAŻ

	<1 rok	1-2 lata	3+ lat
zdacydowanie się nie zgadzam	5	5	4
nie zgadzam się	6	10	1
nie mam zdania	8	26	6
zgadzam się	19	75	6
zdacydowanie się zgadzam	3	24	2

Najbardziej pozytywne odpowiedzi (zgadzam się i zdecydowanie się zgadzam) pochodzą od osób z doświadczeniem 1-2 lata, co sugeruje, że osoby na tym etapie kariery widzą największą wartość badanego zagadnienia. Osoby z najmniejszym stażem (<1 rok) są bardziej podzielone, częściej nie mają zdania lub udzielają odpowiedzi negatywnych. Osoby z największym stażem (3+ lata) rzadko wybierają odpowiedzi skrajne, co może oznaczać, że ich ocena sytuacji jest bardziej neutralna.

Pracownicy z 1-2 latami stażu są najbardziej pozytywnie nastawieni do badanego zagadnienia. Osoby z krótszym stażem mogą wymagać większej ilości informacji lub wsparcia, aby mogły bardziej świadomie ocenić sytuację.

Osoby na stanowiskach kierowniczych częściej wybierają opcję "zgadzam się" i "zdecydowanie się zgadzam", co może świadczyć o ich większym zadowoleniu. Osoby niepełniące funkcji kierowniczych są bardziej podzielone – widoczny jest większy odsetek neutralnych i negatywnych odpowiedzi.

Pracownicy na stanowiskach kierowniczych mają bardziej pozytywne podejście do badanego aspektu, natomiast pracownicy bez funkcji kierowniczych mogą czuć się mniej związani

Tabela 11: Tabela wielodzielcza dla zmiennych PYT 1 i CZY_KIER

	TAK	NIE
zdacydowanie się nie zgadzam	10	4
nie zgadzam się	14	3
nie mam zdania	34	6
zgadzam się	88	12
zdacydowanie się zgadzam	27	2

z tematem lub mieć inne doświadczenia.

Tabela 12: Tabela wielodzielcza dla zmiennych PYT 1 i PŁEĆ

	Kobieta	Mężczyzna
zdacydowanie się nie zgadzam	3	11
nie zgadzam się	7	10
nie mam zdania	14	26
zgadzam się	36	64
zdacydowanie się zgadzam	11	18

Mężczyźni są bardziej podzieleni, częściej wybierają opcje neutralne i negatywne. Kobiety częściej wybierają pozytywne odpowiedzi, co może świadczyć o lepszym dopasowaniu badanego zagadnienia do ich oczekiwań.

Istnieją różnice w postrzeganiu badanego tematu w zależności od płci – warto byłoby zbadać, co wpływa na większą satysfakcję kobiet.

Tabela 13: Tabela wielodzielcza dla zmiennych PYT 1 i WIEK KAT

	0-35	36-45	46-55	55+
zdacydowanie się nie zgadzam	1	11	2	0
nie zgadzam się	6	7	1	3
nie mam zdania	3	24	5	8
zgadzam się	13	50	25	12
zdacydowanie się zgadzam	3	12	12	2

Najbardziej pozytywne odpowiedzi pochodzą od osób w wieku 36-45 lat oraz 46-55 lat – te grupy częściej wybierają opcje “zgadzam się” i “zdecydowanie się zgadzam”. Najmłodsza grupa (0-35 lat) oraz najstarsza grupa (55+) mają wyższy odsetek odpowiedzi neutralnych lub negatywnych.

Pracownicy w średnim wieku są najbardziej pozytywnie nastawieni do badanego zagadnienia. Osoby młodsze i starsze mogą mieć inne oczekiwania lub mniej doświadczenia w tym obszarze.

Zadanie 6

Sporządzono tablicę wielodzielczą dla pary zmiennych: PYT_2 i PYT_3 (po kilku tygodniach ponowna odpowiedź na pytanie dotyczące wsparcia w rozwoju zawodowym i możliwości awansu w firmie).

Tabela 14: Tabela wielodzielcza dla zmiennych PYT 1 i PYT 2

	zdecyd. nie zgadz.	nie zgadz.	zgadz.	zdecyd. zgadzam
zdecyd. nie zgadz.	13	0	1	0
nie zgadz.	16	0	0	1
nie mam zdania	39	0	1	0
zgadz.	3	17	0	80
zdecyd. się zgadzam	3	3	0	23

Dane obejmują zarówno zmienne ilościowe (np. STAŻ, WIEK, PYT_1), jak i jakościowe (np. DZIAŁ, PŁEĆ, CZY_KIER). Brak wartości brakujących sugeruje, że zestaw danych jest kompletny i dobrze przygotowany do analizy. Najwięcej osób w próbie należy do działu PD, najmniej do IT. Większość badanych ma staż pracy 1-2 lata, co może wskazywać na młodą kadrę pracowników. Większość respondentów pełni funkcję kierowniczą. Przewaga mężczyzn w próbie. Największa grupa wiekowa to 36-45 lat.

PYT_1 a DZIAŁ: Najwięcej pozytywnych odpowiedzi ("zgadzam się" lub "zdecydowanie się zgadzam") udzieliły osoby z działu PD.

PYT_1 a STAŻ: Wśród osób z najkrótszym stażem (<1 rok) więcej było odpowiedzi neutralnych lub negatywnych. Może to wynikać z braku doświadczenia lub innego spojrzenia na temat badania.

PYT_1 a CZY_KIER: Kierownicy częściej zgadzali się z twierdzeniem zawartym w PYT_1 niż osoby niepełniące funkcji kierowniczych.

PYT_1 a PŁEĆ: Mężczyźni częściej wyrażali zdecydowaną opinię niż kobiety, które częściej wybierały opcję "nie mam zdania".

PYT_1 a WIEK_KAT: Najwięcej osób w wieku 36-45 lat zgadzało się z PYT_1, natomiast najmniej zdecydowanych odpowiedzi było w grupie 0-35 lat.

PYT_1 a PYT_2: Istnieje silna zależność – osoby, które nie zgadzały się z PYT_1, często miały podobne odpowiedzi na PYT_2, a ci, którzy zgadzali się z PYT_1, także zgadzali się z PYT_2.

Występują różnice w opiniach w zależności od działu, stażu pracy, płci, funkcji kierowniczej i wieku. Kierownicy oraz osoby z dłuższym stażem mają bardziej zdecydowane opinie. Istnieje silna korelacja między odpowiedziami na PYT_1 i PYT_2. Możliwe różnice w podejściu do badanej kwestii między kobietami a mężczyznami oraz młodszymi a starszymi pracownikami.

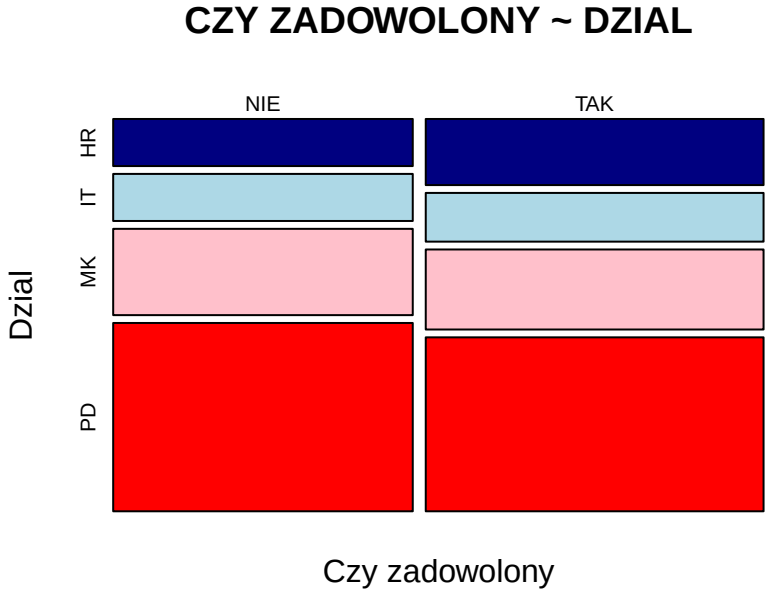
Zadanie 7

Utworzono zmienną CZY_ZADOW na podstawie zmiennej PYT_2 łącząc kategorie "nie zgadzam się" i "zdecydowanie się nie zgadzam" oraz "zgadzam się" i "zdecydowanie się zgadzam"

```
df$CZY_ZADOW <- cut(df$PYT_2,
breaks = c(-3, 0, 2),
labels = c('NIE', 'TAK'))
```

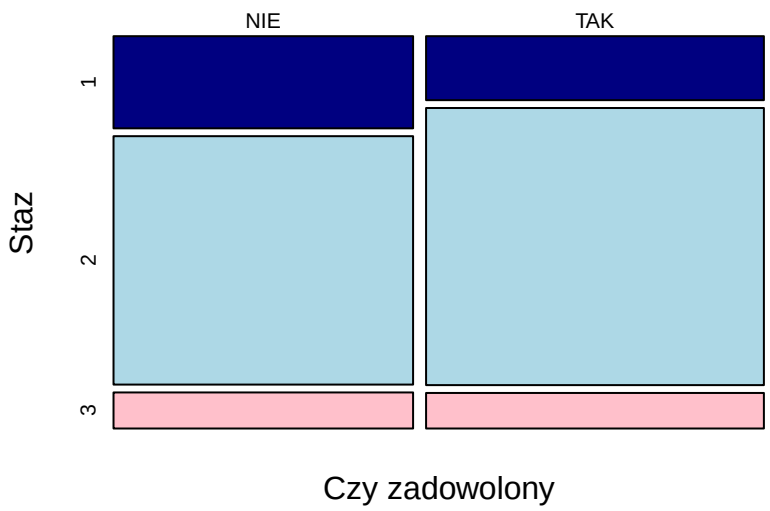
Zadanie 8

Sporządzono wykresy mozaikowe odpowiadające parom zmiennych CZY_ZADOW i DZIAŁ, CZY_ZADOW i STAŻ, CZY_ZADOW i CZY_KIER, CZY_ZADOW i PŁEĆ oraz CZY_ZADOW i WIEK_KAT.



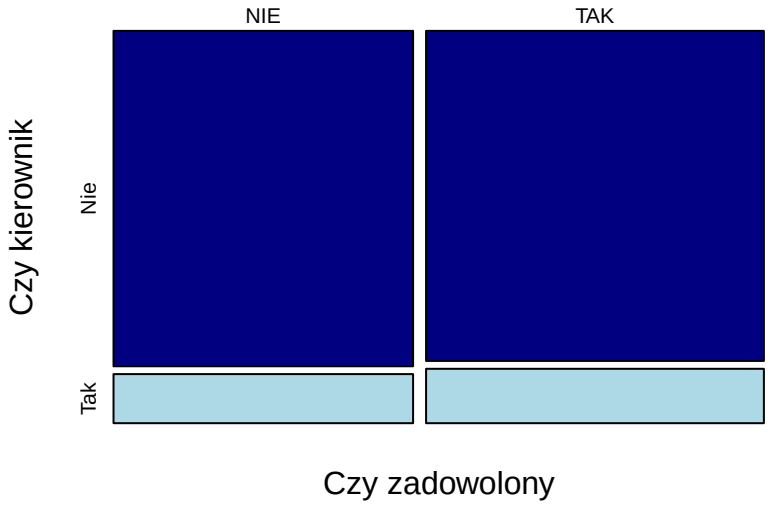

```
## NULL
```

Zadowolenie nie ma związku z działem, do którego należy pracownik. W każdym dziale jest tyle samo zadowolonych, jak i niezadowoloncy pracowników.



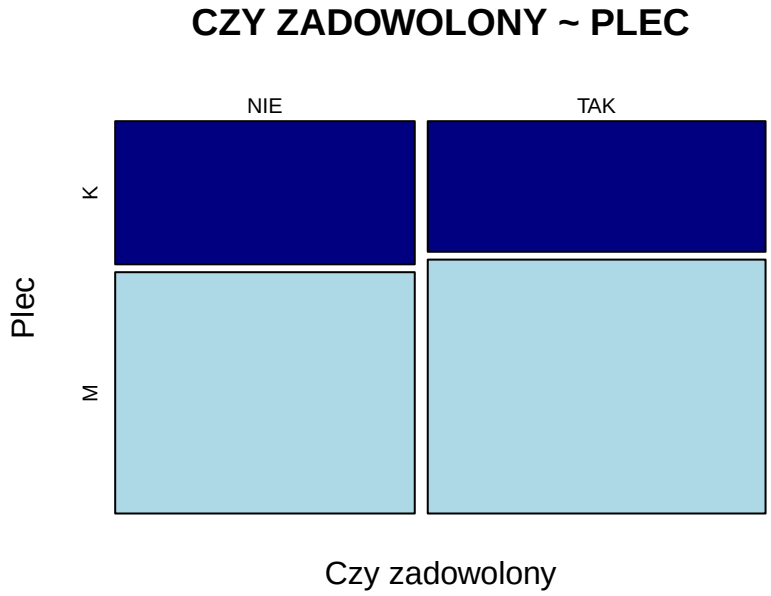

```
## NULL
```

Im pracownik ma krótszy staż w pracy (1-2 lat), tym batrdziej możliwe, że odpowie twierdząco na pytanie. Może być to związane z mniejszą znajomością polityki firmy i większą nadzieją na awans na początku kariery.



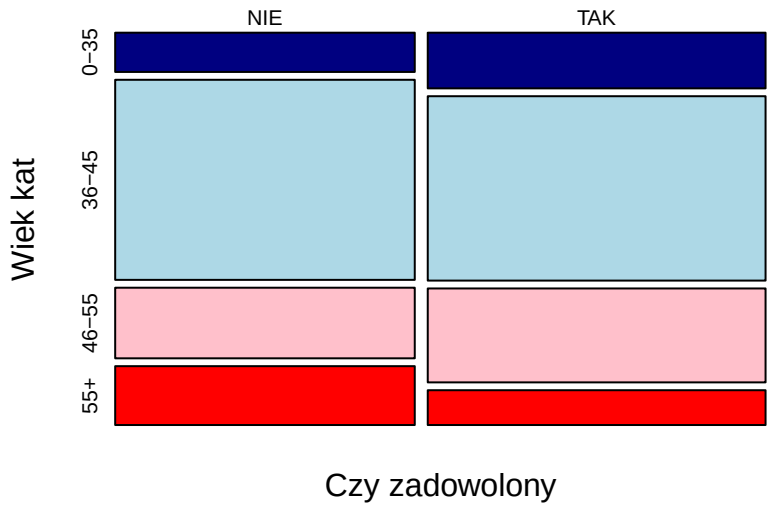

```
## NULL
```

Nie ma znaczenia stanowisko kierownicze, na zadowolenie pracownika.




```
## NULL
```

Mężczyźni są częściej zadowoleni z oferowanych szkoleń i wsparcia niż kobiety. Może być to spowodowane częstymi różnicami w zarobkach kobiet i mężczyzn na tych samym stanowiskach.



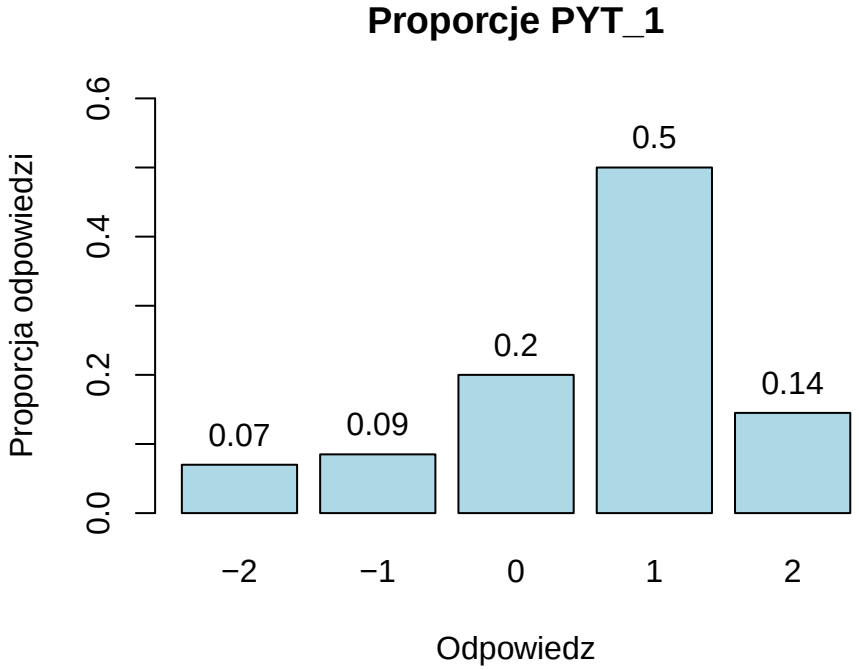

```
## NULL
```

Najmłodzi pracownicy są częściej zadowoleni z pracy, jednak z wiekiem się to zmienia. Pracownicy po 45. roku życia rzadziej są zadowoleni z opcji awansu oferowanych przez firmę, co jest wypadkową zdobytego doświadczenia oraz znajomością (być może słabej) organizacji firmy.

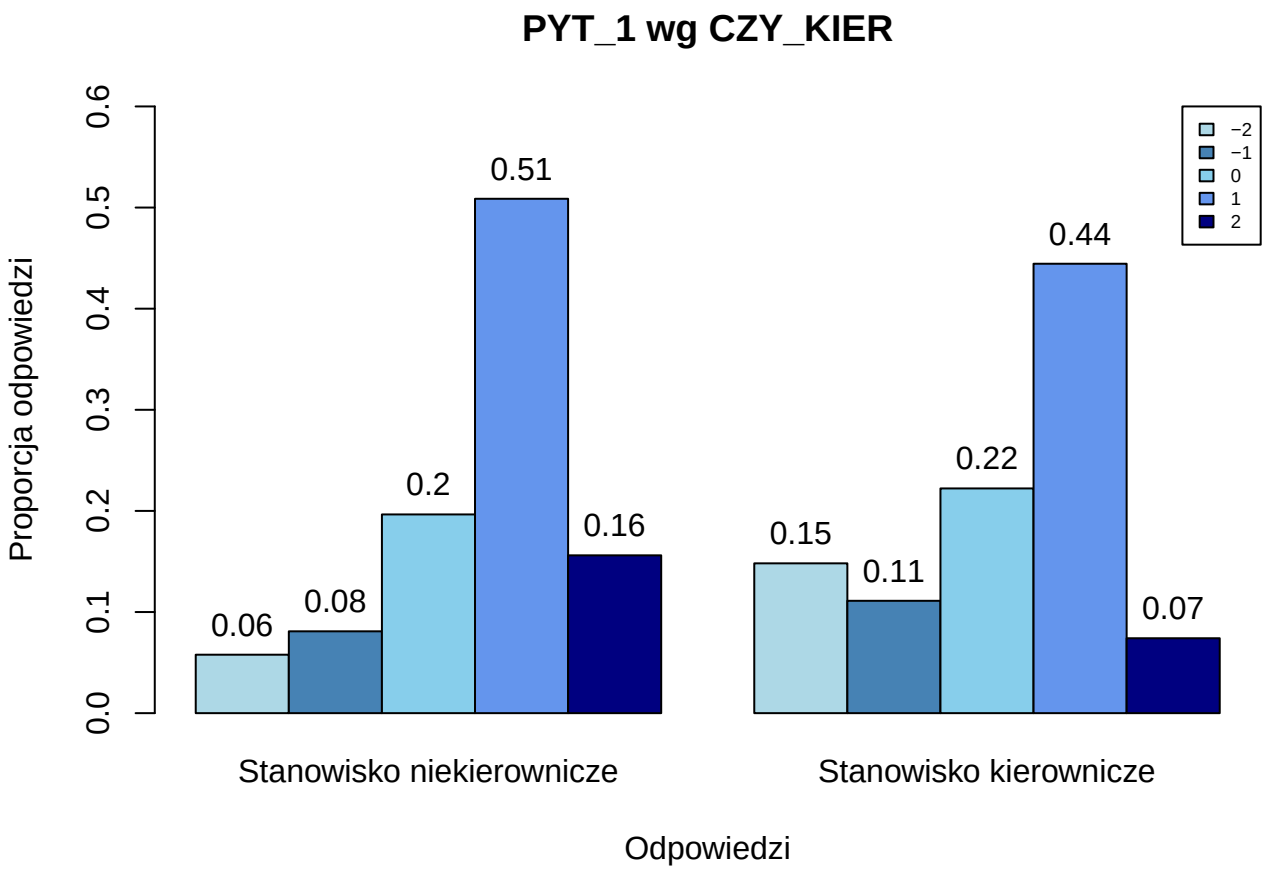
Część II

Zadanie 2

Na wykresie słupkowym została przedstawiona proporcja odpiedzi pracowników firmy na pytanie *PYT_1: "Jak bardzo zgadzasz się ze stwierdzeniem, że firma zapewnia odpowiednie wsparcie i materiały umożliwiaja¸ce skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń?"*. Z wykresu wynika, że połowa badanych zgadza się ze stwierdzeniem, że firma zapewnia odpowiednie wsparcie i materialy umożliwiające skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń, ponadto 14% badanych zdecydowanie popiera tę tezę, a 20% nie ma zdania na ten temat. Natomiast niecałe 10% nie zgadza się z powyższym stwierdzeniem, a 7% uważa, że jest ono zdecydowanie sprzeczne.



Ponadto sprawdzono jak rozkłada się poziom zgodności z powyższym stwierdzeniem względem pełnionego stanowiska (kierownicze lub niekierownicze) dzięki zmiennej CZY_KIER. Z wykresu można wywyniksować, że w obu przypadkach około połowy badanych zgadza się ze stwierdzeniem z PYT_1, jednak ponad dwókrotnie większa część osób (procentowo) bez stanowiska kierowniczego niż na stanowisku kierowniczym jest zdecydowanie zadowolona z udostępnianych materiałów ze szkoleń. Również można zauważyć, że odpowiedzi *nie mam zdania/ nie zgadzam się/ zdecydowanie się nie zgadzam* zanaczyło większy odetek osób na stanowistach kierowniczych niż nie. Zatem z analizy wykresu pudełkowego wynika, że pracownicy na stanowiskach kierowniczych są mniej zdowoleni ze wsparcia i materiałów zapewnianych przez firmę umożliwiaja¸ce skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń.



Zadanie 3

Funkcja *sample()* z biblioteki *stats* losuje próbkę z podanego zbioru danych. **Składnia:** sample(x, size, replace, prob) Gdzie:

- *x* - wektor do losowania
- *size* - liczba elementów do wylosowania
- *replace* - określa czy losowanie jest ze zwracaniem (TRUE/FALSE)
- *prob* - prawdopodobieństwa dla poszczególnych elementów(parametr opcjonalny).

```
# Losowanie 5 liczb z zakresu 1-10
sample(1:10, size = 5)
```

```
## [1] 4 3 2 8 1
```

```
# Losowanie 5 liczb z powtórzeniami
sample(1:10, size = 5, replace = TRUE)
```

```
## [1]  1  5 10  9  9
```

```
# Losowanie z różnymi prawdopodobieństwami
sample(1:10, size = 5, prob = c(0.1, 0.05, 0.15, 0.1, 0.2, 0.05, 0.05, 0.1, 0.1, 0.1))
```

```
## [1] 4 1 3 5 6
```

Następnie z rekordów zawartych w pliku ankieta.csv zostało wylosowane 10% losowych ze wszystkich rekordów za pomocą losowania ze zwracaniem oraz bez zwracania. **Losowanie wierszy ze zwracaniem**

```
##     DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK WIEK_KAT CZY_ZADOW
## 98     PD    2      Nie     0    -2    -2    M   40    36-45       NIE
## 58     PD    2      Nie     1     2     2    M   53    46-55       TAK
## 71     PD    2      Nie     1     2     1    M   28     0-35       TAK
## 196    HR    2      Nie     1     2     2    M   42    36-45       TAK
## 13     IT    2      Tak     1     2     2    K   48    46-55       TAK
## 64     PD    2      Nie     2    -1    -1    M   53    46-55       NIE
## 79     PD    2      Nie    -1    -2    -2    K   38    36-45       NIE
## 195    HR    3      Tak     1     2    -1    M   26     0-35       TAK
## 188    HR    2      Nie     1     2     2    M   48    46-55       TAK
## 45     PD    1      Nie     1     2     2    M   36    36-45       TAK
## 19     IT    2      Nie     1     2    -1    K   34     0-35       TAK
## 169    MK    2      Nie     2     2     2    M   38    36-45       TAK
## 143    MK    2      Nie     1    -1     1    K   52    46-55       NIE
## 36     PD    1      Nie    -2    -2     1    M   29     0-35       NIE
## 10     IT    2      Nie     2    -1     1    K   47    46-55       NIE
## 120    PD    2      Nie     1     2     2    M   38    36-45       TAK
## 23     IT    2      Nie    -1    -2    -2    K   60      55+       NIE
## 68     PD    1      Nie     1     2     2    M   28     0-35       TAK
## 114    PD    1      Nie    -1    -2    -2    M   44    36-45       NIE
## 105    PD    2      Nie     1     2    -1    K   37    36-45       TAK
```

Losowanie wierszy bez zwracania

```
##     DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK WIEK_KAT CZY_ZADOW
## 153    MK    2      Nie     1    -1    -1    M   65      55+       NIE
## 120    PD    2      Nie     1     2     2    M   38    36-45       TAK
## 147    MK    3      Nie     2     2     2    K   46    46-55       TAK
## 105    PD    2      Nie     1     2    -1    K   37    36-45       TAK
## 116    PD    1      Nie     1    -1    -1    M   37    36-45       NIE
## 35     PD    1      Nie    -1    -2    -2    M   28     0-35       NIE
## 10     IT    2      Nie     2    -1     1    K   47    46-55       NIE
## 88     PD    1      Nie     1     2     2    M   49    46-55       TAK
## 171    HR    3      Nie    -2    -2    -2    M   49    46-55       NIE
## 63     PD    2      Nie     2     2     2    M   50    46-55       TAK
## 66     PD    2      Nie     1    -1     1    M   62      55+       NIE
## 131    MK    3      Nie     0    -2    -2    K   45    36-45       NIE
## 126    MK    2      Nie     1     2     2    K   36    36-45       TAK
## 98     PD    2      Nie     0    -2    -2    M   40    36-45       NIE
## 54     PD    1      Nie     0    -2    -2    M   41    36-45       NIE
## 95     PD    2      Nie     1     2     1    M   36    36-45       TAK
```

```
## 125    MK    2      Nie      1     2     2    K   40    36-45       TAK
## 152    MK    1      Nie      0    -2    -1    M   67      55+       NIE
## 17     IT    2      Nie      0    -2    -2    K   45    36-45       NIE
## 197    HR    2      Nie      1    -1    -1    K   35     0-35       NIE
```

Zadanie 4

Funkcja *binomial_sim* dla każdej próbki generuje 1 lub 0 z prawdopodobieństwem p i zwraca realizację próby. Funckja przyjmuje dwa parametry:

- n - długość próby,
- p - teoretyczne prawdopodobieństwo sukcesu.

Funkcja *binomial_N_sim* wykonuje N prób Monte Carlo z rozkładu dwumianowego korzystając z funkcji *binomial_sim*, zwraca wketor realizacji prób z rozkładu dwumianowego.

Funkcja przyjmuje trzy paramatry:

- n - długość próby
- p - teoretyczne prawdopodobieństwo sukcesu
- N - liczbę prób Monte Carlo

```r
# Funkcja do generowania pojedynczej realizacji rozkładu dwumianowego
binomial_sim <- function(n, p) {
  result <- sum(runif(n) < p)
  return(result)
}

# Funkcja do generowania N realizacji rozkładu dwumianowego
binomial_N_sim <- function(n, p, N) {
  W <- numeric(N)
  for (i in 1:N) {
    W[i] <- binomial_sim(n, p)
  }
  return(W)
}
```

W celu przetestowania poprawności zaproponowanych funkcji, przeprowadzono symulację, której celem było wygenerowanie wektora zmiennych losowych i obliczenie ich charakterystyk, a następnie porównanie wyników empirycznych z teoretycznymi.

Parametry symulacji:

- $n = 100$
- $p = 0.5$
- $N = 10000$

Charakterystyki rozkładu,które zostały uwzględnione:

1. **Średnia rozkładu:**

- **Empiryczna:** obliczona za pomocą funkcji *mean()* na wygenerowanych danych,
- **Teoretyczna:** obliczona na podstawie wzoru np, gdzie n to liczba prób, p to prawdopodobieństwo sukcesu.

2.**Odchylenie standardowe rozkładu:**

- **Empiryczne:** obliczone za pomocą funkcji *sd()* na wygenerowanych danych,
- **Teroretyczne:** obliczone na podstawie wzoru

$$\sigma = \sqrt{np(1-p)}$$

, gdzie n to liczba prób, p to prawdopodobieństwo sukcesu.

3. **Histogram częstości:**

- **Empiryczne:** przedstawione za pomocą funkcji *hist()* na podstawie wygenerowanych danych
- **Teoretyczne:** porównane z teoretycznymi wartościami prawdopodobieństwa sukcesu, obliczonymi za pomocą funkcji *dbinom()*.

Wyniki symulacji sugerują, że funkcja poprawnie generuje zmienną losową z rozkładu dwumianowego.

```
# Parametry rozkładu dwumianowego
n <- 100
p <- 0.5
N <- 10000

# Generowanie danych
simulated_data <- binomial_N_sim(n, p, N)

# Teoretyczne wartości
theoretical_mean <- n * p
theoretical_sd <- sqrt(n * p * (1 - p))

# Empiryczne wartości
empirical_mean <- mean(simulated_data)
empirical_sd <- sd(simulated_data)

cat("Teoretyczna średnia: ", theoretical_mean, "\n")
```

```
## Teoretyczna średnia:  50
```

```r
cat("Teoretyczne odchylenie standardowe: ", theoretical_sd, "\n")
```

```
## Teoretyczne odchylenie standardowe:  5
```

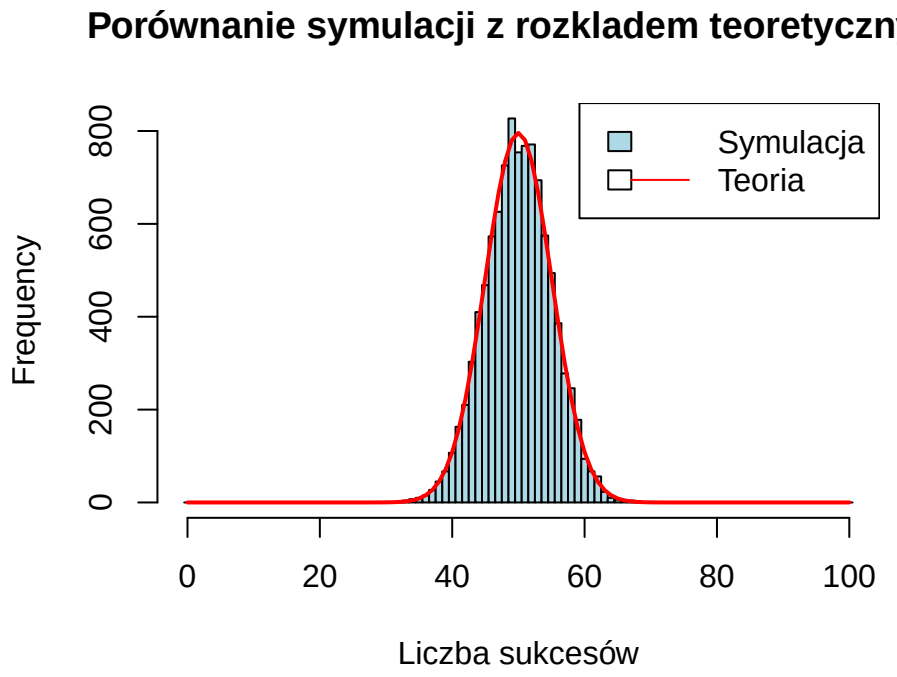
```r
cat("Empiryczna średnia: ", empirical_mean, "\n")
```

```
## Empiryczna średnia:  50.0212
```

```r
cat("Empiryczne odchylenie standardowe: ", empirical_sd, "\n")
```

```
## Empiryczne odchylenie standardowe:  5.003964
```

```r
# Teoretyczne prawdopodobieństwa rozkładu dwumianowego
x_vals <- 0:n
theoretical_probs <- dbinom(x_vals, size = n, prob = p) * N

# Histogram wyników symulacji
hist(simulated_data,
     breaks = seq(-0.5, n + 0.5, by = 1),
     main = "Porównanie symulacji z rozkładem teoretycznym",
     xlab = "Liczba sukcesów",
     col = "lightblue",
     border = "black",
     freq = TRUE,
     right = FALSE)

# Dodanie teoretycznych wartości
lines(x_vals, theoretical_probs, col = "red", lwd = 2)  # Linia
legend("topright", legend = c("Symulacja", "Teoria"),
       fill = c("lightblue", NA), col = c("black", "red"),
       lty = c(NA, 1), border = "black")
```



Zadanie 5

Funkcja *wielomianowy_sim* generuje pojedynczą realizację rozkładu wielomianowego z prawdopodbieństwami sukcesu danymi jako wektor p.

Przyjmuje dwa parametry:

- n - liczba prób (wielkość próby)
- p - wektor prawdopodobieństw dla poszczególnych kategorii.

Zwraca wektor liczności wystąpień każdej kategorii.

Funkcja *wielomianowy_N_sim* wykonuje N prób Monte Carlo dla rozkładu wielomianowego. Korzysta z funkcji *wielomianowy_sim*, aby wygenerować realizacje prób i zwraca macierz wyników. Przyjmuje trzy parametry:

- n - liczba prób (wielkość próby)
- p - wektor prawdopodobieństw dla poszczególnych kategorii
- N - liczba prób Monte Carlo

```r
# Funkcja generująca pojedynczą realizację rozkładu wielomianowego
wielomianowy_sim <- function(n, p) {
  k <- length(p)
  proby <- sample(1:k, size = n, replace = TRUE, prob = p)
  tab <- table(factor(proby, levels = 1:k))
  as.vector(tab)
}
```

```r
# Funkcja wykonująca N prób Monte Carlo dla rozkładu wielomianowego
wielomianowy_N_sim <- function(n, p, N) {
  k <- length(p)
  W <- replicate(N, wielomianowy_sim(n, p))
  result <- W
  return(result)
}
```

Analogicznie do poprzedniego zadania, przeprowadzono symulację, której celem było sprawdzenie poprawności zaproponowanych funckji. Została wygenerowana macierz zawierająca realizaje zmiennych rozkładu wielomianowego, a następnie obliczono ich charakterystyki i porównano z teoretycznymi wartościami.

Parametry symulacji:

- n = 100
- p = [0.5, 0.1, 0.2, 0.2]
- N = 10000

Charakterystyki rozkładu,które zostały uwzględnione:

1. **Średnia rozkładu:**

- **Empiryczna:** obliczona za pomocą funkcji *rowMeans()* na wygenerowanych danych,
- **Teoretyczna:** obliczona na podstawie wzoru np, gdzie n to liczba prób, p to prawdopodobieństwo sukcesu.

2.**Odchylenie standardowe rozkładu:**

- **Empiryczne:** obliczone za pomocą funkcji *apply(simulated_data, 1, sd)* na wygenerowanych danych,
- **Teroretyczne:** obliczone na podstawie wzoru

$$\sigma = \sqrt{np(1-p)}$$

, gdzie n to liczba prób, p to wektor prawdopodobińst w.

Wyniki symulacji sugerują, że funkcja poprawnie generuje zmienną losową z rozkładu wielomianowego.

```
## Teoretyczna średnia:  50 10 20 20
```

```
## Teoretyczne odchylenie standardowe:  5 3 4 4
```

```
## Empiryczna średnia:  49.9856 10.0324 19.9026 20.0794
```

```
## Empiryczne odchylenie standardowe:  5.02859 3.006169 4.011848 3.961548
```

Część III i IV

Zadanie 6

Funkcja *clopper_pearson_ci* oblicza przedział ufności dla proporcji na podstawie metody Cloppera-Pearsona, która jest dokładnym podejściem do wyznaczania przedziału ufności dla proporcji w przypadku prób losowych. Jest to podejście, które jest szczególnie użyteczne w przypadku małych prób lub rzadkich wydarzeń, gdzie standardne metody mogą nie być dokładne.

Funkcja przyjmuje parametry:

- *confidence* - poziom ufności
- *successes* - liczba sukcesów
- *trials* - liczba prób
- *data* - opcjonalny wektor danych binarnych (0 lub 1).

Jeśli data jest podane, funkcja automatycznie wylicza liczbę sukcesów *successes* i liczbę prób *trials* na podstawie danych. Funkcja zwraca dolną i górna granicę przedziału ufności, korzystając z wbudowanej funkcji *qbeta*.

```
clopper_pearson_ci <- function(
    confidence,
    successes = NULL,
    trials = NULL,
    data = NULL) {
  if (!is.null(data)) {
    successes <- sum(data)
    trials <- length(data)
  }
  if (is.null(successes) || is.null(trials)) {
    stop("Należy podać liczbę sukcesów i prób lub wektor danych.")
  }
  alpha <- 1 - confidence
  lower_bound <- qbeta(alpha / 2, successes, trials - successes + 1)
  upper_bound <- qbeta(1 - alpha / 2, successes + 1, trials - successes)
  return(c(lower_bound, upper_bound))
}

# Przykład użycia
confidence_level <- 0.95
successes <- 30
trials <- 100
ci <- clopper_pearson_ci(confidence_level,
                         successes = successes,
```

```
                          trials = trials)
cat("Przedział ufności (Clopper-Pearson) dla",
    successes, "/", trials, ":", ci, "\n")
```

```
## Przedział ufności (Clopper-Pearson) dla 30 / 100 : 0.2124064 0.3998147
```

```
# Generowanie losowych danych binarnych
data <- rbinom(100, 1, 0.3)
ci_data <- clopper_pearson_ci(confidence_level, data = data)
cat("Przedział ufności (Clopper-Pearson)
    dla danych binarnych:", ci_data, "\n")
```

```
## Przedział ufności (Clopper-Pearson)
##     dla danych binarnych: 0.2302199 0.4207669
```

Zadanie 7

Na podstawie zmiennej *PYT_3* (które jest tym samem pytaniem co PYT_2, tylko zadane kilka tygodni później) stworzono zmienną *CZY_ZADOW_2*.

```
df$CZY_ZADOW_2 <- cut(df$PYT_3, breaks = c(-3, 0, 2),
                      labels = c("NIE", "TAK"),
                      include.lowest = TRUE)

successes_1 <- sum(df$CZY_ZADOW == "TAK", na.rm = TRUE)
trials_1 <- sum(df$CZY_ZADOW %in% c("TAK", "NIE"), na.rm = TRUE)
ci1<- clopper_pearson_ci(0.95,successes_1,trials_1)


successes_2 <- sum(df$CZY_ZADOW_2 == "TAK", na.rm = TRUE)
trials_2 <- sum(df$CZY_ZADOW_2 %in% c("TAK", "NIE"), na.rm = TRUE)
ci2<- clopper_pearson_ci(0.95,successes_2,trials_2)

cat("Przedział ufności (Clopper-Pearson) dla CZY_ZADOW_2:", ci2,"\n",
    "Przedział ufności (Clopper-Pearson) dla CZY_ZADOW:", ci1, "\n")
```

```
## Przedział ufności (Clopper-Pearson) dla CZY_ZADOW_2: 0.5184216 0.6588694
##  Przedział ufności (Clopper-Pearson) dla CZY_ZADOW: 0.4583305 0.6007671
```

Przedziały ufności dla prawopodobieństwa sukcesu (u nas odpowiedź "TAK"), po kilku tygodniach zmieniły się. Po czasie pracownicy częściej wyrażali zadowolenie. Możliwe, że firma po przeanalizowaniu wcześniejszych wyników ankiety, postanowiła wprowadzić zmiany, które wpływają pozytywnie na pracowników.

Zadanie 8

Funkcja wbudowana *rbinom* generuje dane binarne(0/1) na podstawie rozkładu dwumianowego. Przyjmuje parametry:

- *n* - liczba prób,
- *size* - liczba obserwacji w każdej próbie,
- *p* - prawdopodobieństwo sukcesu.

Wynikiem jest wektor *losowe_wartosci*, który zawiera liczbę sukcesów (liczba 1) w każdej z 1000 prób (o długości 10).

Funkcja *binom.confint* oblicza przedział ufności dla proporcji sukcesów w próbach binarnych. Argument *conf.level = 0.95* ustawia poziom ufności na 95%. Argument *methods = "exact"* oznacza, że do obliczenia przedziału ufności zostanie użyta dokładna metoda Cloppera-Pearsona. Argument *methods = asymptotic"* oznacza, że do obliczenia przedziału ufności zostanie użyta dokładna metoda Walda. Argument *methods = wilson"* oznacza, że do obliczenia przedziału ufności zostanie użyta dokładna metoda Wilsona.

```
n <- 1000      # liczba prób
p <- 0.5     # prawdopodobieństwo sukcesu
size <- 10 # liczba generowanych wartości


losowe_wartosci <- rbinom(n, size, p)
sukcesy <- sum(losowe_wartosci)
proby <- length(losowe_wartosci)

ci_cp <- binom.confint(sukcesy, n = proby*size,
                       methods = "exact",
                       conf.level = 0.95)
cat("Przedział ufności Cloppera-Pearsona:", ci_cp$lower, ci_cp$upper, "\n")
```

```
## Przedział ufności Cloppera-Pearsona: 0.4940505 0.5137472
```

```
ci_wald <- binom.confint(sukcesy, n = proby*size,
                         conf.level = 0.95,
                         methods="asymptotic",
                         correct = FALSE)
cat("Przedział ufności Wald'a:", ci_wald$lower, ci_wald$upper, "\n")
```

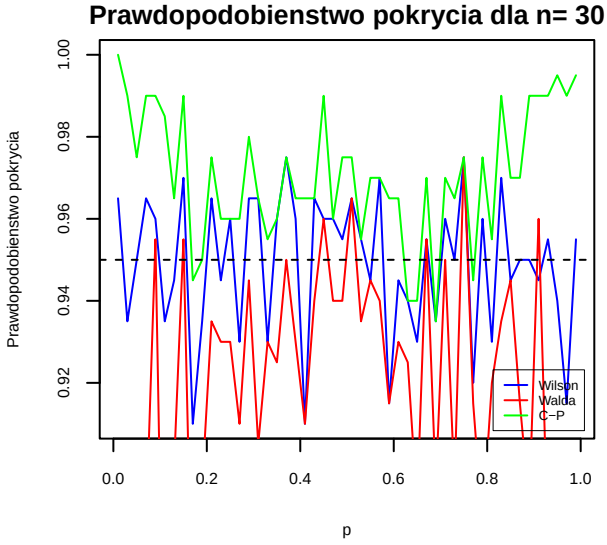
```
## Przedział ufności Wald'a: 0.4941005 0.5136995
```

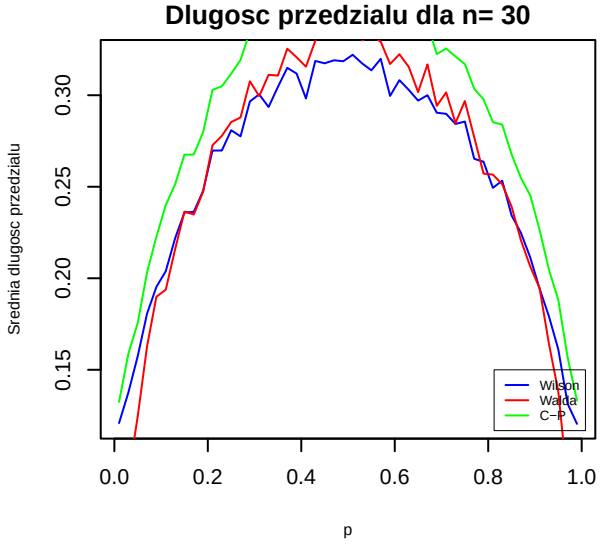
```
ci_wilson <- binom.confint(sukcesy, n = proby*size,
                           methods = "wilson",
                           conf.level = 0.95)
cat("Przedział ufności Wilsona:", ci_wilson$lower, ci_wilson$upper, "\n")
```

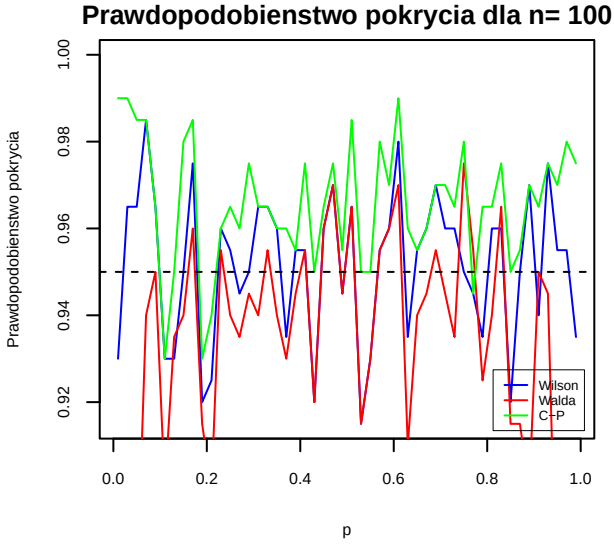
```
## Przedział ufności Wilsona: 0.4941009 0.5136961
```

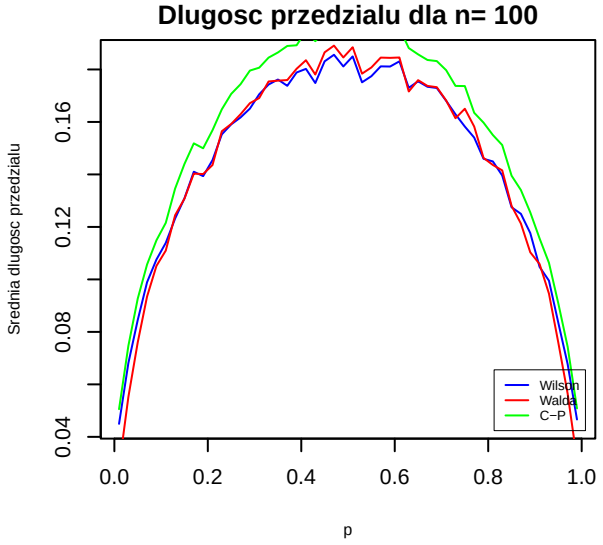
Metody Wald'a oraz Wilsona są do siebie najbardziej zbliżone. Metoda Clopper-Pearsona jest dokładna, ale często daje szerokie przedziały ufności.

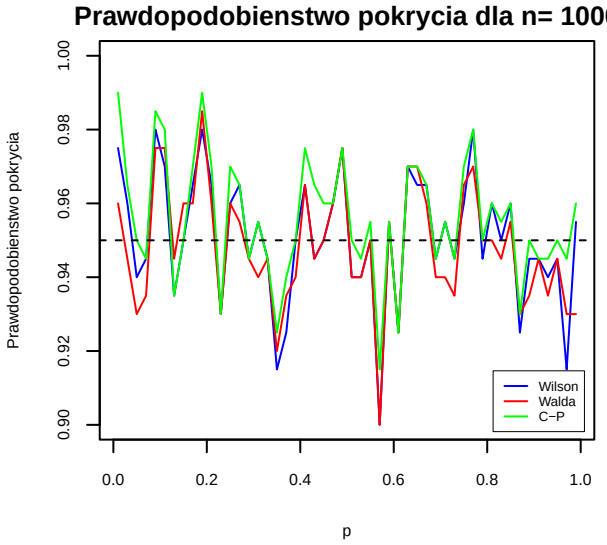
Zadanie 9

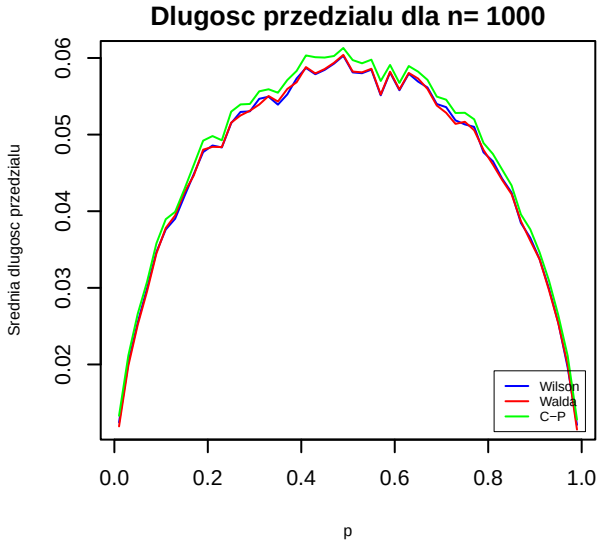

Prawdopodobienstwo pokrycia dla n= 30
Prawdopodobienstwo pokrycia
p
Wilson
Walda
C−P


Dlugosc przedzialu dla n= 30
Srednia dlugosc przedzialu
p
Wilson
Walda
C−P


Prawdopodobienstwo pokrycia dla n= 100
Prawdopodobienstwo pokrycia
p
Wilson
Walda
C−P


Dlugosc przedzialu dla n= 100
Srednia dlugosc przedzialu
p
Wilson
Walda
C−P


Prawdopodobienstwo pokrycia dla n= 1000
Prawdopodobienstwo pokrycia
p
Wilson
Walda
C−P


Dlugosc przedzialu dla n= 1000
Srednia dlugosc przedzialu
p
Wilson
Walda
C−P

Dla małych próbek (n=30) metoda Wald'a wykazuje dużą niestabilność, z wyraźnymi spadkami poniżej poziomu pozostałych metod. W miarę wzrostu liczności próby (n=100, n=1000) różnice między metodami się zmniejszają, a wszystkie trzy metody zbliżają się do oczekiwanego poziomu pokrycia.

Na wykresach widać, istotną wadę przedziału Cloppera-Pearsona tj. są "ekstremalnie konserwatywne", ponieważ prawdopodobieństwo pokrycia nieznanego parametru p jest większe od poziomu ufności 95% dla wszytskich p. Dzieje się tak ze względu na ich długość. Faktycznie na wykresach długości przedziałów, metoda C-P osiąga największe wartości wśród metod, które wraz z zwiększającą się wartościa n, zmiejszają się. Na wykresach prawdopodobieństwa pokrycia wykres metody C-P ma swoje minimum w okolicach p=0.5. Natomiast metody Wilsona i Wald'a mają tam swoje maksimum. Wraz z większającą sie liczbą n wszystkie 3 metody wypłaszczają się w okolicy prawdopodobieństwa pokrycia 0.95.

Wszystkie metody wykazują podobny kształt – długość przedziału jest największa dla p=0.5 i maleje dla wartości bliższych 0 i 1. Dla n=30 metoda Wald'a daje krótsze przedziały, ale kosztem gorszegoprawdopodobieństwa pokrycia. Dla większych próbek długości przedziałów uzyskane różnymi metodami są bardzo podobne.

Przy większych n różnice między metodami stają się mniej istotne – wszystkie trzy metody dają podobne prawdopodobieństwo pokrycia i długości przedziałów.Metoda Walda nie sprawdza się dobrze dla małych prób – daje niestabilne pokrycie i krótsze przedziały kosztem wiarygodności. Metody Wilsona i C-P są bardziej niezawodne, szczególnie dla mniejszych wartości n.

Część V

Zadanie 10

1. **Test dokładny** W języku programowania R do wykonywania testu dokładnego służy funckja *binom.test()*, która wykonuje dokładny test dwumianowy (test dokładny dla proporcji) oparty na rozkładzie dwumianowym.\ Stosowany, gdy próbka jest mała, ponieważ nie opiera się na przybliżeniach asymptotycznych. Funkcja przyjmuje argumenty:

- x - liczba sukcesów w próbie
- n - liczba obserwacji w próbie
- p - prawdopodonbieństwo sukcesu
- *alternative* - paramter określający hipotezę typ hipotezy alternatywnej: `"two.sided"` (dwustronna, domyślna), `"greater"` (większa) lub `"less"` (mniejsza)
- *conf.level* - poziom ufności dla testu (domyślnie = 0.95)

2. **Test asymptotyczny** Do wykonywania testu asyptotycznego język programownaia R wykorzystuje funkcję *prop.test()*, który jest testem z użyciem asymptotycznego przybliżenia normalnego (test chi-kwadrat dla proporcji). Stosowany dla dużych prób, ponieważ opiera się na centralnym twierdzeniu granicznym. W celu sprawdzenia parametrów podanych funkcji przerowadzono symaulację z pramatrami:

- x - liczba sukcesów w próbie
- n - liczba obserwacji w próbie
- p - prawdopodonbieństwo sukcesu
- *alternative* - paramter określający hipotezę typ hipotezy alternatywnej: `"two.sided"` (dwustronna, domyślna), `"greater"` (większa) lub `"less"` (mniejsza)
- *conf.level* - poziom ufności dla testu (domyślnie = 0.95)
- *correct* - Yates' continuity correction to poprawka stosowana w teście chi-kwadrat, szczególnie w przypadku małych prób, mająca na celu poprawienie dokładności testu. Zmienia statystykę testową, aby zredukować nadmierne wartości p w małych próbach, ponieważ w takich przypadkach rozkład chi-kwadrat może być mniej dokładny.

1. Symulacja I:

- $x1 = 3$
- $n1 = 10$
- $p1 = 0.5$

```
## Wynik testu dokładnego:
```

```
## P-value: 0.34375
```

```
## Wynik testu alternatywnego:
```

```
## P-value: 0.3427817
```

```
## Poziom: 0.05
```

Według wyników testów wynika, że nie ma podstwa do odrzucenia H0, ponieważ p-value jest większe niż poziom istotności.

2. Symulacja II:

- $x2 = 260$
- $n2 = 500$
- $p2 = 0.5$

```
## Wynik testu dokładnego:
```

```
## P-value: 0.3955108
```

```
## Wynik testu alternatywnego:
```

```
## P-value: 0.3954887
```

```
## Poziom: 0.05
```

Według wyników testów, nie ma podstaw do odrzucenia hipotezy zerowej (H), ponieważ p-value jest większe niż poziom istotności.

Różnica w wynikach p-value dla testu dokładnego (binom.test) i testu proporcji (prop.test) wynika z różnych metod obliczeń:

- *binom.test* to test dokładny, który używa rozkładu dwumianowego i jest bardziej precyzyjny w przypadku małych prób.

- *prop.test* to test aproksymacyjny, który bazuje na rozkładzie normalnym, co jest przybliżeniem i może prowadzić do niewielkich różnic w p-value, szczególnie w małych próbach.

W tym przypadku, wyniki p-value są bardzo podobne, ale test binominalny jest dokładniejszy. Warto zauważyć, że dla większych prób (symulacja II) wyniki p-value są bliższe sobie, co sugeruje, że test aproksymacyjny staje się bardziej dokładny w miarę wzrostu rozmiaru próby.

Zadanie 11

Za pomocą funkcji *binom.test()* oraz *prop.test()* przerprowadzono weryfikację pięciu hipotez na poziomie ufności $1- \ = 0.95$.

Zadanie 11.1

Pierwsza hipoteza dotyczyła jaki jest stosunek pracujących w firmie kobiet. Postawiona hipoteza statystyczna:

$H_0 : p = 0.5$

Wyniki testów przedstawione w tabelce poniżej są znacząco mniejsze od ustalonego poziomu istotności, zatem odrzucamy hipotezę zerową, co sugeruje, że udział kobiet w firmie nie jest równy 50%

```
##         Test      P_value Interpretation
## 1 binom.test 4.972973e-05   Odrzucamy H0
## 2  prop.test 5.565628e-05   Odrzucamy H0
```

Zadanie 11.2

Druga hipoteza dotyczy pozytywnej odpowiedzi na drugie pytanie zadane w ankiecie. Postawiona hipoteza statystyczna:

$H_0 : p >= 0.7$

Ta hipoteza również została odrzucona, ponieważ wyniki testów przedstawione poniżej są drastycznie mniejsze niż ustalony poziom istotności, po oznacza, że mniej niż 70% badanych uważa szkolenia za przystosowane do ich potrzeb.

```
##         Test      P_value Interpretation
## 1 binom.test 3.212877e-07   Odrzucamy H0
## 2  prop.test 1.175729e-07   Odrzucamy H0
```

Zadanie 11.3

Kolejna hipoteza dotyczy porónania czy prawdopodobieństwo, że kobieta pracuje na stanowisku kierowniczym jest równe prawdopodobieństwu, że mężczyzna pracuje na stanowisku

kierowniczym. Do zweryfikowania tej hipotezy wykorzstano kolumnę CZY_KIER określającą czy pracownik jest na stanowisku kierowniczym ("Tak") czy nie ("Nie"). Postawiona hipoteza statystyczna:

$$H_0 : p = \frac{\text{liczba mężczyzn na stanowisku kierowniczym}}{\text{liczba wszystkich pracujących mężczyzn}}$$

W tym przypadku wyniki obu testów przedstawionych w poniższej tabeli wyszły większe niż ustalony poziom istotności, a zatem nie ma podstaw na odrzucenie postawione hipotezy zerowej. Oznacza to, że możliwe, że w firmie procentowo jest tyle samo kobiet i mężczyzn na stanowisku kierowniczym.

```
##         Test   P_value                Interpretation
## 1 binom.test 0.5040621 Brak podstaw do odrzucenia H0
## 2  prop.test 0.5121591 Brak podstaw do odrzucenia H0
```

Zadanie 11.4

Przedostatnia hipotezy związana jest z optymistyczną odpowiedzią na drugie pytanie z ankiety względem płci. Sprawdzamy czy równy procent kobiet i mężczyzn uważa, że szkolenia są przystosowane do ich potrzeb. Postawiona hipoteza statystyczna:

$$H_0 : p = \frac{\text{liczba mężczyzn zadowolonych ze szkolenia}}{\text{liczba wszystkich pracujących mężczyzn}}$$

Z przerowadzonych testó wynika, że nie ma podstaw do odrzucenia hipotezy zerowej. Oznacza to, że możliwe, że w firmie pocentowo tyle samo kobiet i mężczyzn odpowiedziało pozytywnie na drugie pytanie.

```
##         Test   P_value                Interpretation
## 1 binom.test 0.5538817 Brak podstaw do odrzucenia H0
## 2  prop.test 0.6291586 Brak podstaw do odrzucenia H0
```

Zadanie 11.5

Ostatnia postawiona hipoteza sprawdza czy prawdopodbieństwo, że losowa kobieta z firmy pracuje w dziale zasobów ludzkich jest większe lub równe niż procent mężczyzn pracujących w tym dziale. Postawiona hipoteza statystyczna:

$$H_0 : p = \frac{\text{liczba mężczyzn pracujących w dziale HR}}{\text{liczba wszystkich pracujących mężczyzn}}$$

Według wyników przedstawionych w poniższej tabelce nie ma podstaw do odrzucenia hipotezy zerowej, a zatem nie można określić czy procentowo więcej lub tyle samo kobiet pracuje w dziel HR niż mężczyzn.

```
##         Test      P_value Interpretation
## 1 binom.test 0.0003456683   Odrzucamy H0
## 2  prop.test 0.0012537251   Odrzucamy H0
```

Przeprowadzona analiza pozwoliła zweryfikować założone hipotezy oraz określić, czy istnieją istotne statystycznie różnice w badanych aspektach struktury zatrudnienia i opinii pracowników. Uzyskane wyniki mogą być podstawą do dalszych analiz oraz ewentualnych działań w zakresie polityki zatrudnienia i organizacji szkoleń w firmie.

Zadanie 12

Do wyznaczenia mocy testu binomalnego (dokładnego) oraz mocy testu proporcji (asymptotycznego) przeprowadzono analizę dla hipotezy zerowej: $H_0 : p = 0.9$ przeciwko hipotezie alternatywnej $H_1 : p \neq 0.9$ na poziomie ufności 0.95.

Moce testów zostały wyznaczone dla następujących wartości parametru $n = (50, 100, 500, 1000)$.

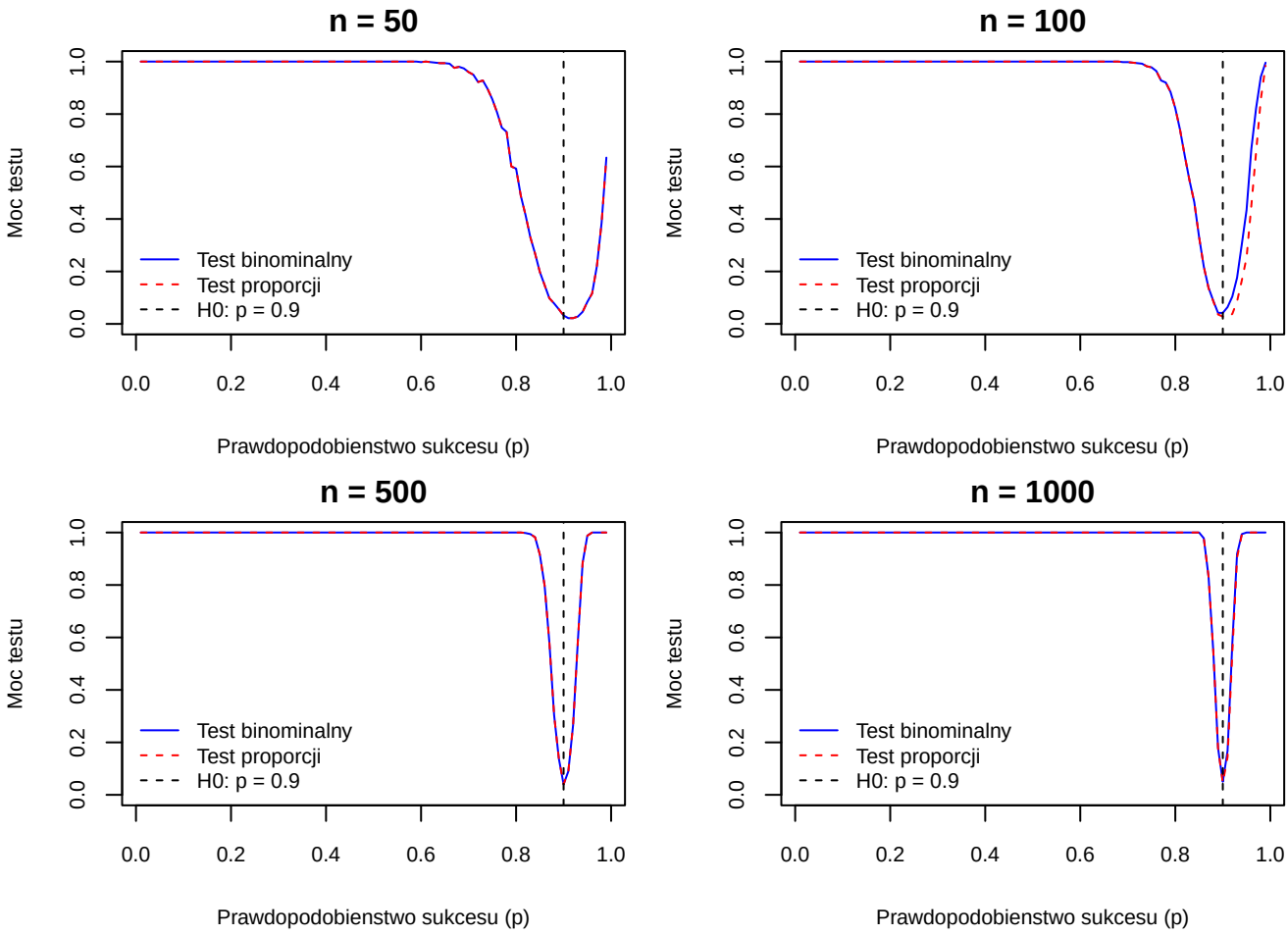
Na poniższych wykresach przedstawiono zależność mocy testu od wartości parametru $p \in [0, 1]$. Można zauważyć, że dla większości wartości p funkcja mocy osiąga wartości bliskie 1, jednak w okolicach $p = 0.9$ występuje wyraźny spadek.

Z poniższych wykresów można wywnioskować wpływ liczności próby na wyniki testów jest również istotny. Dla mniejszych prób, na przykład przy $n = 50$ lub $n = 100$, spadek mocy w pobliżu $p = 0.9$ jest łagodniejszy. Testy mają trudność w rozróżnieniu wartości bliskich hipotezie zerowej, co prowadzi do mniejszej precyzji w wykrywaniu różnic. Z kolei dla większych prób, takich jak $n = 500$ lub $n = 1000$, spadek mocy staje się bardziej stromy. Oznacza to, że testy z większą licznością prób wykrywają nawet niewielkie odchylenia od $p = 0.9$ z większą precyzją.

```r
# Funkcja powertest
powertest <- function(N, n, alpha, p0) {
  p_values <- seq(0.01, 0.99, by = 0.01)
  l <- length(p_values)
  binom_vec <- numeric(l)
  prop_vec <- numeric(l)

  for (j in 1:l) {
    p <- p_values[j]
    binom_vec_N <- numeric(N)
    prop_vec_N <- numeric(N)

    for (i in 1:N) {
      binom <- rbinom(1, n, p)
      binom_test <- binom.test(binom, n, p0, conf.level = 1 - alpha)
      binom_vec_N[i] <- binom_test$p.value <= alpha

      prop_test <- prop.test(binom, n, p0, conf.level = 1 - alpha)
```

```
      prop_vec_N[i] <- prop_test$p.value <= alpha
    }

    binom_vec[j] <- mean(binom_vec_N)
    prop_vec[j] <- mean(prop_vec_N)
  }

  return(list(power_bin = binom_vec, power_prop = prop_vec))
}
```

```
N <- 500
n_values <- c(50, 100, 500, 1000)
alpha <- 0.05
p0 <- 0.9

results <- list()

for (i in 1:length(n_values)) {
  n <- n_values[i]
  results[[i]] <- powertest(N, n, alpha, p0)
}
```



Podsumowując, wzrost liczności próby poprawia precyzję testów. Dla dużych prób, funkcja mocy gwałtownie maleje wokół $p = 0.9$, co sprawia, że testy są w stanie lepiej odróżnić wartości zgodne z hipotezą zerową od tych, które się od niej różnią. Choć test binomialny i test proporcji dają podobne wyniki, test proporcji może być mniej dokładny dla małych prób, dlatego w praktyce dla małych próbek zaleca się stosowanie testu binomialnego. Dla większych prób test proporcji może być wystarczająco precyzyjny i stanowić bardziej efektywne rozwiązanie.

Zadanie dodatkowe

Funkcja logit - przekształcenie stosowane do prawdopodobieństwa, które przekształca wartości z przedziału $(0, 1)$ na zakres $(\infty, -\infty)$, o wzorze:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

gdzie:

- p - prawdopodobieństwo sukcesu w eksperymencie.

Funkcja logit jest używana w regresji logistycznej, gdzie stosuje się ją do przekształcenia proporcji na skalę, która może przyjmować dowolne wartości rzeczywiste. Dzięki temu możemy modelować zależności pomiędzy zmiennymi niezależnymi a prawdopodobieństwem wystąpienia określonego zdarzenia.

Metoda delta (znana również jako metoda przybliżenia wariancji funkcji lub metoda liniowej aproksymacji) - polega na zastosowaniu liniowej aproksymacji do funkcji estymatora, aby obliczyć jego wariancję lub inne miary niepewności. Stosowana, gdy estymatorem jest funkcja trudna do bezpośredniego obliczenia jednak możliwe jest przybliżenie liniowe wokół punktu, którym jest średnia tego estymatora.

W poniższym kodzie została zimplementowana funkcja *granice_asymptotycznego_przedziału_ufności()*, która ma na celu obliczenie asymptotycznego przedziału ufności prawdopodobieństwa sukcesu p, wykorzystując funkcję logit oraz metodę delta. Przyjmuje parametry:

- n - wielkość próby
- p - prawdopodobieństwo sukcesu
- *alpha* - poziom istotności

```
granice_asymptotycznego_przedziału_ufności <- function(n,p, alpha){
  x <- rbinom(1, n, p)
  p_est = x/n #estymacja proporcji sukcesów

  #Przekształcenie logit
  p_logit_est = log(p_est/(1 - p_est))

  #Metoda delta do obliczenia wariancji przekształconego estymatora
  var_p_logit_est = (1 / (p_est * (1 - p_est))) * (1 / n)

  kwantyl_alpha <- qnorm(1 - alpha)

  #Granice przedziału ufności na skali logit
  logit_lower = p_logit_est - kwantyl_alpha * sqrt(var_p_logit_est)
logit_upper = p_logit_est + kwantyl_alpha * sqrt(var_p_logit_est)

  #Odwracanie funkcji ligit
  p_lower = 1 / (1 + exp(-logit_lower))
  p_upper = 1 / (1 + exp(-logit_upper))


  result <- c(p_lower, p_upper)
  return(result)
}
```

W celu sprawdzenia poprawności zaimplementowanej funkcji przeprowadzono sumulację porównującą obliczony asymptotyczny przedział ufności prawdopodobieństwa sukcesu p

przez tę funkcję i porównano ją z wynikiem obliczonym przy pomocy funkcji wbudowanej *prop.test()* korzytając z argumentu *conf.int*.

Parametry symulacji:

- $n = 1000$
- $p = 0.7$
- $alpha = 0.05$

Wyniki obliczone za pomocą obu funkcji co sugeruje poprawność zaimplementowanej metody do znajdowania granic asyptotycznego przedziału ufności dla prawdopodobieństwa sukcesu p.

```
## [1] "Granice asymptotycznego przedziału ufności (nasza metoda):"
```

```
## [1] 0.669520 0.717432
```

```
## [1] "Granice asymptotycznego przedziału ufności (prop.test()):"
```

```
## [1] 0.6703646 0.7280778
## attr(,"conf.level")
## [1] 0.95
```