Sprawozdanie 1

Joanna Kusy, Tomasz Srebniak

Część I

Zadanie 1

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

W ankiecie zostały umieszczone odpowiedzi na poniższe pytania:

- W jakim dziale pracujesz?" - zmienna **DZIAŁ** przyjmująca wartości: **HR** (Dział zasobów ludzkich), **IT** (Dział technologii informatycznych), **PD** (Dział Produktowy) lub **MK** (Dział Marketingu),
- "Jak długo pracujesz w firmie?" - zmienna **STAŻ** przyjmująca wartości: **1** (Poniżej jednego roku), **2** (Między jednym a trzema latami) lub **3** (Powyżej trzech lat),
- "Czy pełnisz funkcję kierowniczą?" - zmienna **CZY_KIER** przyjmująca wartości: **Tak** (Stanowisko kierownicze) lub **Nie** (Stanowisko inne niż kierownicze),
- "Jak bardzo zgadzasz sie ze stwierdzeniem, że firma zapewnia odpowiednie wsparcie i materiały umożliwiające skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń?" - zmienna **PYT_1** przyjmuja ca wartości: **-2** (zdecydowanie się nie zgadzam), **-1** (nie zgadzam się), **0** (nie mam zdania), **1** (zgadzam się), **2** (zdecydowanie się zgadzam).
- "Jak bardzo zgadzasz się ze stwierdzeniem, że firma oferuje szkolenia dostosowane do twoich potrzeb, wspierając twój rozwój zawodowy i szanse na awans?" - zmienna **PYT_2** przyjmująca wartości: **-2** (zdecydowanie się nie zgadzam), **-1** (nie zgadzam się), **1** (zgadzam się), **2** (zdecydowanie się zgadzam).

Dodatkowo w ramach metryczki ankietowani zostali poproszeni o wskazanie swojego wieku - zmienna **WIEK** przyjmująca wartości numeryczne, oraz wskazanie płci - zmienna **PŁEĆ** przyjmująca wartość **K** lub **M**.

Kilka tygodni później w firmie przeprowadzono cykl szkoleń indywidualnie dostosowanych do potrzeb konkretnych grup pracowników. Ankietowanych biorących udział w badaniu poproszono wówczas o ponowną odpowiedź na pytanie dotyczące wsparcia w rozwoju zawodowym i możliwości awansu w firmie - zmienna **PYT_3**.

Podpunkt 1

Poniżej przedstawiono pierwsze pięć rekordów wyników ankiety, których nie poddano żadnych modyfikacjom.

```
ankieta <- read.csv('ankieta.csv', header = TRUE, sep = ";", check.names = F)
head(ankieta, 5)
```

```
  DZIA\xa3 STA\xaf CZY_KIER PYT_1 PYT_2 PYT_3 P\xa3E\xc6 WIEK
1       IT       2      Nie     1    -2     1          M   64
2       IT       2      Nie     0    -2    -2          M   67
3       IT       2      Nie     1     2     2          M   65
4       IT       2      Nie    -1    -2    -2          K   68
5       IT       3      Tak     1     2    -1          K   65
```

Nazwy kolumn zawierające polskie znaki nie wczytały się prawidłowo. W celu zapewnienia czytelności danych poprawiono nazwy kolumn, w których występowały błędy.

```
colnames(ankieta)[1] <- "DZIAŁ"
colnames(ankieta)[2] <- "STAŻ"
colnames(ankieta)[7] <- "PŁEĆ"
```

Poniżej przedstawiono pierwsze pięć rekordów wyników ankiety z poprawionymi nazwami kolumn.

```
  DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK
1    IT    2      Nie     1    -2     1    M   64
2    IT    2      Nie     0    -2    -2    M   67
3    IT    2      Nie     1     2     2    M   65
4    IT    2      Nie    -1    -2    -2    K   68
5    IT    3      Tak     1     2    -1    K   65
```

Zadbano także o odpowiednie typy badanych zmiennych, czego efekty zaprezentowano poniżej.

```r
sapply(ankieta, class)
```

```
    DZIAŁ       STAŻ   CZY_KIER      PYT_1      PYT_2      PYT_3       PŁEĆ       WIEK
 "factor"   "factor"   "factor"   "factor"   "factor"   "factor"   "factor"  "integer"
```

Zbadano także, czy zmienne przyjmują wartości zgodne z zamieszczonym wyżej opisem.

```r
all(unique(ankieta$DZIAŁ) %in% c("HR", "IT", "PD", "MK"))
```

```
[1] TRUE
```

```r
all(unique(ankieta$STAŻ) %in% c(1, 2, 3))
```

```
[1] TRUE
```

```r
all(unique(ankieta$CZY_KIER) %in% c("Tak", "Nie"))
```

```
[1] TRUE
```

```r
all(unique(ankieta$PYT_1) %in% c(-2, -1, 0, 1, 2))
```

```
[1] TRUE
```

```r
all(unique(ankieta$PYT_2) %in% c(-2, -1, 1, 2))
```

```
[1] TRUE
```

```r
all(unique(ankieta$PYT_3) %in% c(-2, -1, 1, 2))
```

```
[1] TRUE
```

```r
all(unique(ankieta$PŁEĆ) %in% c("K", "M"))
```

```
[1] TRUE
```

```
cat("Minimalna wartość dla zmiennej WIEK:", min(ankieta$WIEK),
    ", a maksymalna:", max(ankieta$WIEK), ".")
```

```
Minimalna wartość dla zmiennej WIEK: 25 , a maksymalna: 70 .
```

Wartości we wszystkich kolumnach są zgodne z naszymi oczekiwaniami. Dodatkowo sprawdzono, czy ankieta nie zawiera braków w danych.

```
missing_values <- sum(is.na(ankieta))
missing_values
```

```
[1] 0
```

Ankieta nie zawiera żadnych braków danych.

Podpunkt 2

Utworzono zmienną **WIEK__KAT** przeprowadzając kategoryzację zmiennej **WIEK** korzystając z następujących przedziałów: do 35 lat (młody), między 36 a 45 lat (średni), między 46 a 55 lat (starszy), powyżej 55 lat (wiek przedemerytalny, skrótowo oznaczony jako emerytura).

```
ankieta['WIEK_KAT'] <- ifelse(ankieta$WIEK < 36, 'młody',
                              ifelse(ankieta$WIEK < 46, 'średni',
                                     ifelse(ankieta$WIEK < 56, 'starszy',
                                            'emerytura')))
```

```
  DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK  WIEK_KAT
1    IT    2      Nie     1    -2     1    M   64 emerytura
2    IT    2      Nie     0    -2    -2    M   67 emerytura
3    IT    2      Nie     1     2     2    M   65 emerytura
4    IT    2      Nie    -1    -2    -2    K   68 emerytura
5    IT    3      Tak     1     2    -1    K   65 emerytura
6    IT    3      Tak     0     1     1    K   57 emerytura
```

Podpunkt 3

Tablica liczności dla zmiennej **DZIAŁ**.

```r
ankieta |> group_by(DZIAŁ) |> summarise(n = n())
```

```
# A tibble: 4 x 2
  DZIAŁ     n
  <fct> <int>
1 HR       31
2 IT       26
3 MK       45
4 PD       98
```

Z tabeli wynika, że najwięcej osób pracuje kolejno w dziale: produktowym, marketingu, zasobów ludzkich oraz technologii informatycznych. Sugeruje to, że firma koncentruje się działaniach operacyjnych, czyli związanych z produktami i ich dobrym marketingiem. Działy HR oraz IT pełnią zapewne rolę wspierającą, stąd ich mniejsza liczebność w strukturze firmy.

Tabela liczności dla zmiennej **STAŻ**.

```
# A tibble: 3 x 2
  STAŻ      n
  <fct> <int>
1 1        41
2 2       140
3 3        19
```

Najwięcej pracowników pracuje w fimie od roku do trzech lat. Zapewne jest to okres, w którym nastąpił największy rozwój przedsiębiorstwa. Stosunkowo duża liczba osób zatrudnionych poniżej roku może świadczyć o dalszym rozwoju firmy. Mała liczba pracowników ze stażem dłuższym niż trzy lata może potwierdzać teorię o dynamicznym rozwoju firmy lub świadczyć o problemach z utrzymaniem kadry.

Tabela liczności dla zmiennej **CZY_KIER**.

```
# A tibble: 2 x 2
  CZY_KIER     n
  <fct>    <int>
1 Nie        173
2 Tak         27
```

Z tabeli wynika, że około co piąta osoba zajmuje stanowisko kierownicze. Stosunek liczby kierowników do ogólnej liczby pracowników sugeruje, że firma posiada dobrze rozwiniętą strukturę zarządzania.

Tabela liczności dla zmiennej **PŁEĆ**.

```
# A tibble: 2 x 2
  PŁEĆ      n
  <fct> <int>
1 K        71
2 M       129
```

Z przeprowadzonej ankiety wynika, że w firmie pracuje więcej mężczyzn niż kobiet. Może to świadczyć o dominacji mężczyzn w obszarach związanych z technologią lub wskazywać na potencjalną dyskryminację wobec kobiet podczas procesu rekrutacyjnego.

Tabela liczności dla zmiennej **WIEK_KAT**.

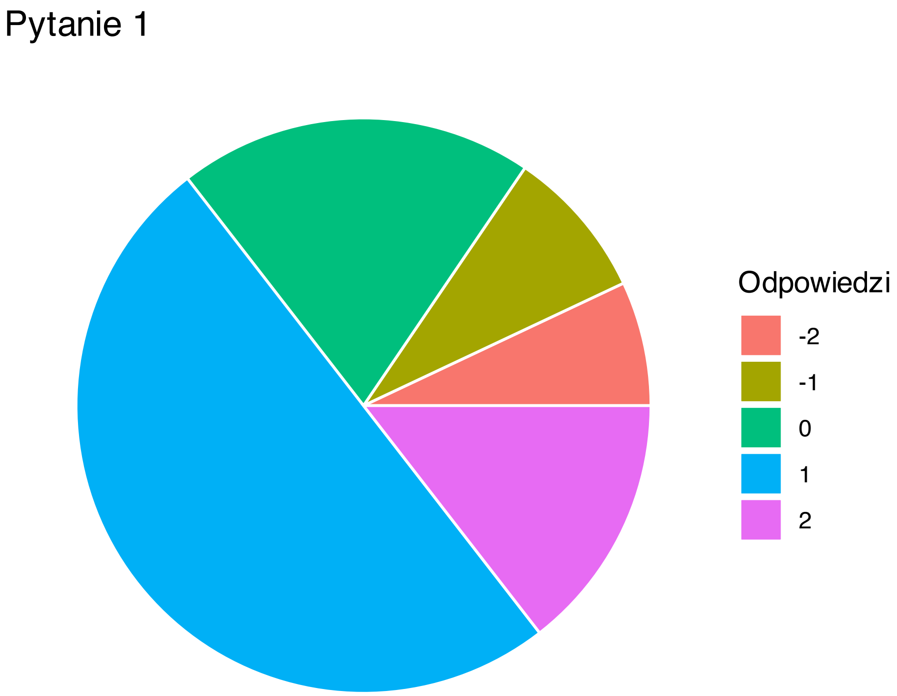
```
# A tibble: 4 x 2
  WIEK_KAT      n
  <chr>     <int>
1 emerytura    25
2 młody        26
3 starszy      45
4 średni      104
```

Wśród ankietowanych osób najwięcej jest między 36. a 55. rokiem życia. Zapewne ludzie w tym wieku posiadają odpowiedni zakres umiejętności i doświadczenie cenione w firmie. Jednocześnie stosunkowo niewielka liczba młodszych pracowników może sugerować ograniczone możliwości dla osób rozpoczynających karierę lub specyfikę branży, która wymaga większej praktyki. Obecność osób w wieku przedemerytalnym wskazuje, że firma ceni wiedzę ekspercką zdobytą dzięki wieloletniemu doświadczeniu. Ich stosunkowo niewielka liczba może wynikać z naturalnej rotacji związanej ze zbliżającą się emeryturą, a także z mniejszej skłonności pracodawcy do inwestowania w rozwój pracowników w tym wieku.

Podpunkt 4

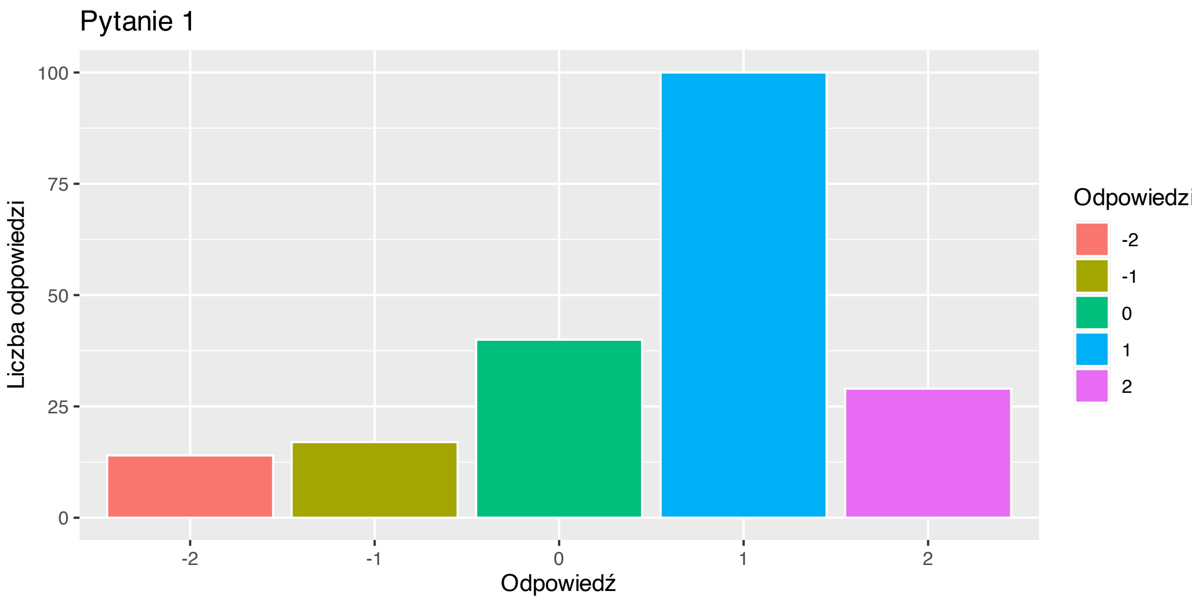
Wykres kołowy dla zmiennej **PYT_1**.

```r
p1 <- ggplot(ankieta, aes(x='', fill=PYT_1)) +
  geom_bar(color='white') +
  coord_polar('y', start=pi/2) +
  theme_void() +
  labs(title='Pytanie 1', fill='Odpowiedzi')
p1
```



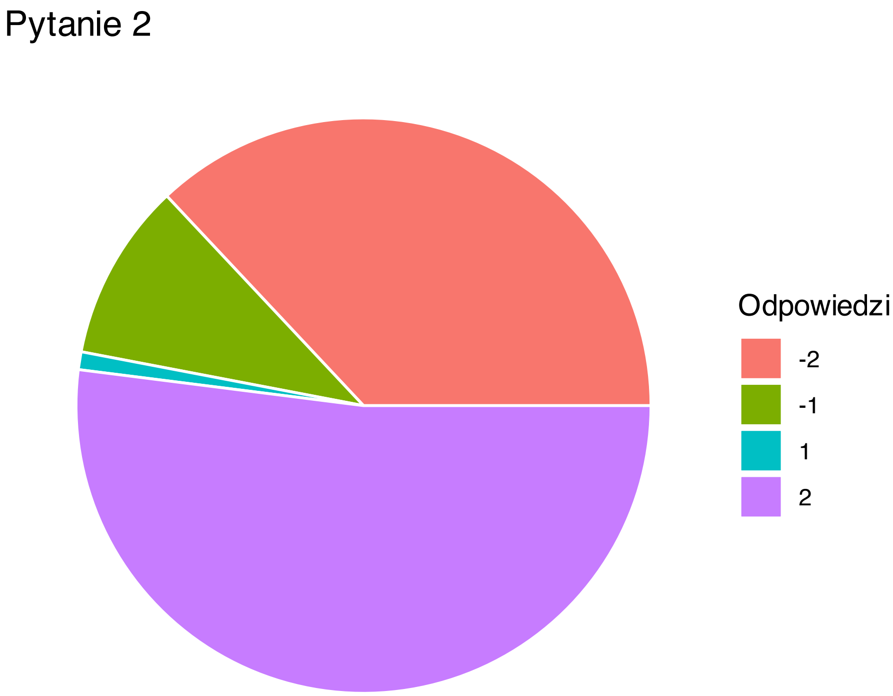
Wykres słupkowy dla zmiennej **PYT_1**.

```
p1 <- ggplot(ankieta, aes(x=PYT_1, fill=PYT_1)) +
  geom_bar(color='white') +
  labs(title='Pytanie 1', fill='Odpowiedzi',
       x='Odpowiedź', y='Liczba odpowiedzi')
p1
```

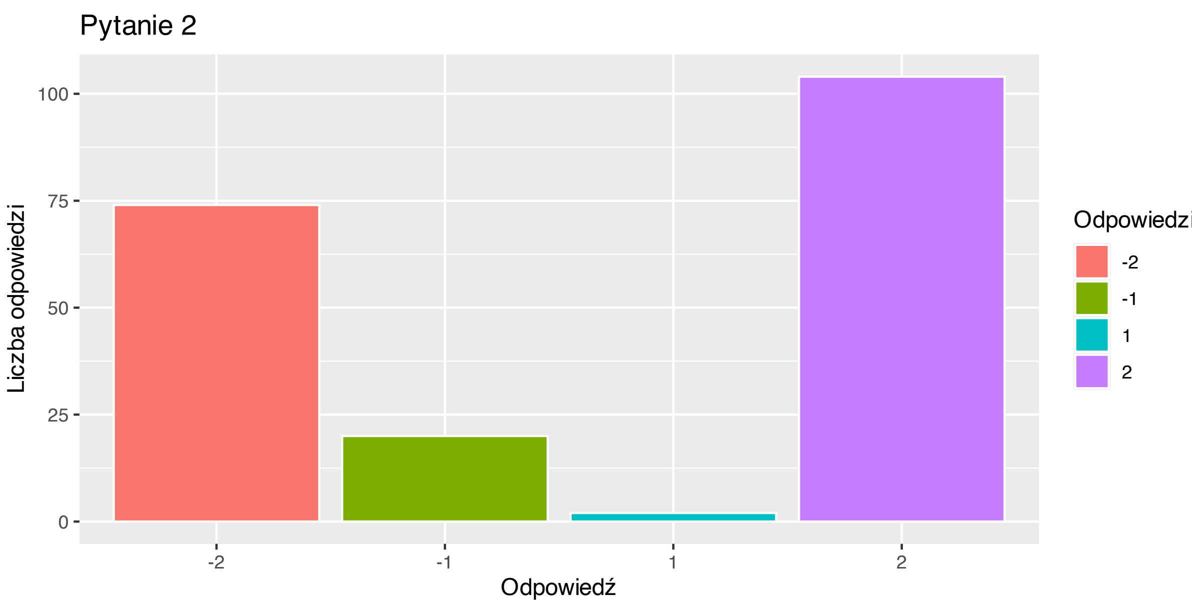


Z wykresów wynika, że ponad połowa badanych zgadza się lub zdecydowanie zgadza się ze stwierdzeniem, iż firma zapewnia odpowiednie wsparcie i materiały umożliwiające skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń. Stosunkowo duża część ankietowanych nie ma zdania w tej kwestii. Mniej niż jedna na cztery badane osoby nie zgadza się lub zdecydowanie nie zgadza się ze stwierdzeniem zawartym w pytaniu. Warto byłoby sprawdzić, czy osoby, które uważają, że firma nie zapewnia im odpowiedniego wsparcia lub materiałów nie są ze sobą powiązane np. przez dział, w którym pracują bądź managera. Może to wskazywać, że sposób zarządzania w niektórych zespołach może nie sprzyjać skutecznemu wykorzystywaniu zdobytej wiedzy i dalszemu rozwojowi pracowników.

Wykres kołowy dla zmiennej **PYT_2**.



Wykres słupkowy dla zmiennej **PYT_2**.



Z wykresów wynika, że badani pracownicy firmy mocno dzielą się w opinii dotyczącej atrakcyjności oferty szkoleń i ich dostosowania do własnych potrzeb. Większość pracowników jest z niej wysoce zadowolona, jednak istnieje także spora wyraźnie nieusatysfakcjonowana grupa osób. Oznacza to, że oferta szkoleniowa skutecznie odpowiada na potrzeby części pracowników, ale jednocześnie nie spełnia oczekiwań innych.

Podpunkt 5

Tablica wielodzielcza dla pary zmiennych **PYT_1** i **DZIAŁ**.

```
ankieta |>
  group_by(DZIAŁ) |>
  summarise(PYT_1 = mean(as.numeric(as.character(PYT_1)), na.rm = TRUE))
```

```
# A tibble: 4 x 2
  DZIAŁ PYT_1
  <fct> <dbl>
1 HR    0.613
2 IT    0.885
3 MK    0.578
4 PD    0.459
```

Powyżej przedstawiono tablicę wielodzielczą dla zmiennej **DZIAŁ** oraz średniej zmiennej **PYT_1**. Wyższa wartość w tabeli oznacza większą satysfakcję z poziomu wsparcia i zapewnienia materiałów umożliwiających skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń. Obecność tych dwóch czynników najlepiej oceniają osoby pracujące w dziale **IT**, nieco gorsze wyniki odnotowano w działach **HR** i **MK**. Swoją sytuację najgorzej ocenią osoby z działu **PD**, co może wskazywać na potrzebę zwiększenia poziomu wsparcia lub lepszego dostosowania dostępu do materiałów wspomagających rozwój pracowników w tym dziale.

Tablica wielodzielcza dla zmiennej **STAŻ** i średniej zmiennej **PYT_1**.

```
# A tibble: 3 x 2
  STAŻ   PYT_1
  <fct>  <dbl>
1 1     0.220
2 2     0.736
3 3     0.0526
```

Grupując poziom satysfakcji na podstawie stażu w firmie, można stwierdzić, że najbardziej zadowoloną grupą są osoby zatrudnione dłużej niż rok, ale krócej niż trzy lata. Być może taki staż pracy pozwala najlepiej docenić poziom wsparcia i dostępność materiałów oferowanych przez pracodawcę. Osoby pracujące krócej niż rok mogły przyjść ze środowisk, w których otrzymywały większą pomoc, lub nie miały jeszcze wystarczająco dużo czasu, aby w pełni ocenić warunki panujące u obecnego pracodawcy. Z kolei mniejsze zadowolenie wśród pracowników z najdłuższym stażem może wynikać z niewielkiej liczby zmian i ulepszeń w systemie wsparcia na przestrzeni lat.

Tablica wielodzielcza dla zmiennej **CZY_KIER** i średniej zmiennej **PYT_1**.

```
# A tibble: 2 x 2
  CZY_KIER PYT_1
  <fct>    <dbl>
1 Nie      0.624
2 Tak      0.185
```

Z tabeli wielodzielczej wynika, że osoby pełniące funkcje kierownicze są mniej zadowolne z poziomu wsparcia oferowanego przez firmę. Może to sugerować, że ich potrzeby w zakresie szkoleń i dostępnych materiałów różnią się od potrzeb pracowników na niższych stanowiskach, a obecny system wsparcia nie jest dostosowany do specyfiki ich roli.

Tablica wielodzielcza dla zmiennej **PŁEĆ** i średniej zmiennej **PYT_1**.

```
# A tibble: 2 x 2
  PŁEĆ  PYT_1
  <fct> <dbl>
1 K     0.634
2 M     0.527
```

Kobiety są bardziej zadowolne z poziomu wsparcia oferowanego przez firmę, jednak różnica pomiędzy płciami jest niewielka i nie wskazuje na znaczące nierówności w dostępie do zasobów czy jakości oferowanej pomocy.

Tablica wielodzielcza dla zmiennej **WIEK_KAT** i średniej zmiennej **PYT_1**.

```
# A tibble: 4 x 2
  WIEK_KAT  PYT_1
  <chr>     <dbl>
1 emerytura 0.52
2 młody     0.423
3 starszy   0.978
4 średni    0.433
```

Wyniki ankiety wskazują, że najbardziej usatysfakcjonowane z pomocy oferowanej przez pracodawcę są osoby między 36. a 45. rokiem życia. Pozostałe grupy wykazują niższy, aczkolwiek zbliżony do siebie poziom satysfakcji. Ze względu na istotną różnicę pomiędzy osobami należącymi do kategorii „starszy" a pozostałymi grupami, warto zbadać, jakie czynniki wpływają na ich odmienną opinię.

Podpunkt 6

Poniżej przedstawiono tabelę wielodzielczą dla pary zmiennych **PYT_2** i **PYT_3**.

```
ankieta |> group_by(PYT_2, PYT_3) |>
  summarise(n = n(), .groups='keep') |>
  pivot_wider(names_from = PYT_3, values_from = n, values_fill = 0)
```

```
# A tibble: 4 x 5
# Groups:   PYT_2 [4]
  PYT_2  `-2`  `-1`   `1`   `2`
  <fct> <int> <int> <int> <int>
1 -2       49    16     5     4
2 -1        3     6    10     1
3 1         0     0     2     0
4 2         0     8    15    81
```

Z tabeli wynika, że szkolenia indywidualnie dostosowane do potrzeb konkretnych grup pracowników nie zmieniły znacząco zdania ankietowanych. Wiele osób o najbardziej skrajnych opiniach pozostało przy swoich początkowych ocenach. Trzydzieści sześć osób zmieniło swoją opinię na lepszą, natomiast dwadzieścia sześć osób – na gorszą.

Podpunkt 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".

```
ankieta['CZY_ZADOW'] <- ifelse(ankieta$PYT_2 %in% c(1, 2),
                               'zadowolony', 'niezadowolony')
head(ankieta)
```

```
  DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK  WIEK_KAT     CZY_ZADOW
1    IT    2      Nie     1    -2     1    M   64 emerytura niezadowolony
2    IT    2      Nie     0    -2    -2    M   67 emerytura niezadowolony
3    IT    2      Nie     1     2     2    M   65 emerytura    zadowolony
4    IT    2      Nie    -1    -2    -2    K   68 emerytura niezadowolony
5    IT    3      Tak     1     2    -1    K   65 emerytura    zadowolony
6    IT    3      Tak     0     1     1    K   57 emerytura    zadowolony
```

Podpunkt 8

Poniżej 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**. Do każdego z nich postawiono hipotezę dotyczącą relacji między zmiennymi.

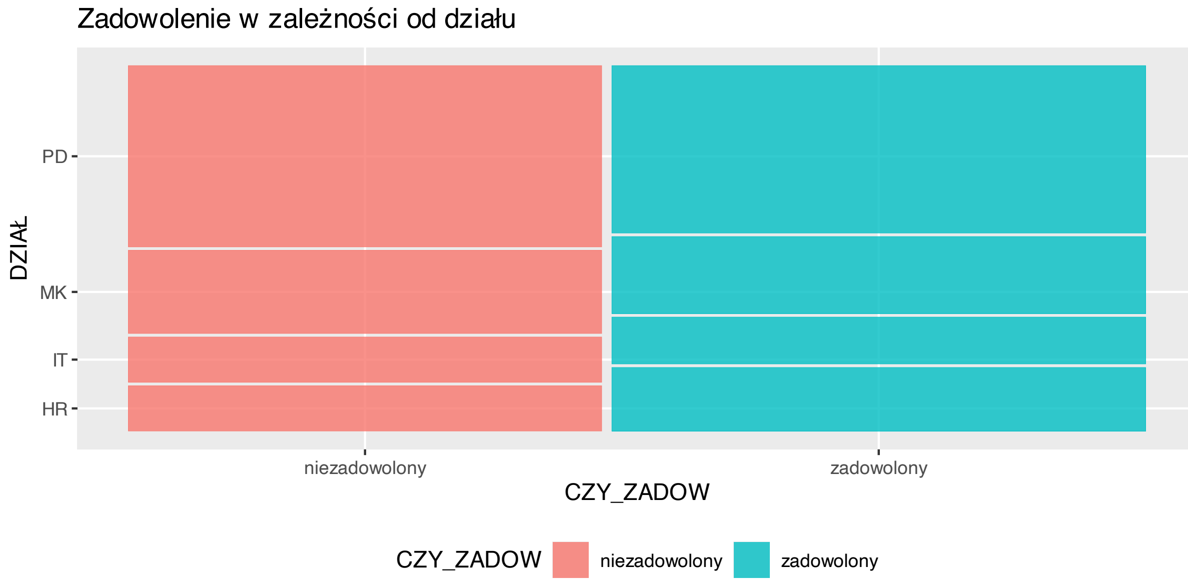
Wykres mozaikowy dla pary zmiennych **CZY_ZADOW** i **DZIAŁ**.

```r
ggplot() +
  geom_mosaic(
    data = ankieta,
    aes(weight = 1, x = product(DZIAŁ, CZY_ZADOW), fill = CZY_ZADOW)
    ) +
  labs(title='Zadowolenie w zależności od działu') +
  theme(legend.position = 'bottom')
```

```
Warning: The `scale_name` argument of `continuous_scale()` is deprecated as of ggplot2
3.5.0.
```

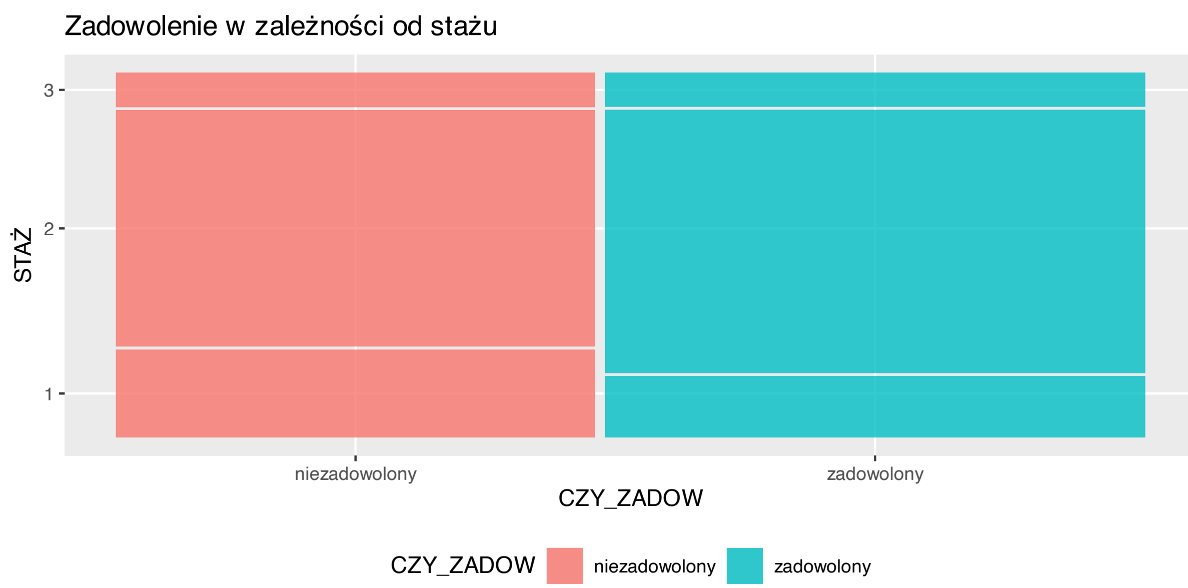
```
Warning: The `trans` argument of `continuous_scale()` is deprecated as of ggplot2 3.5.0.
i Please use the `transform` argument instead.
```

```
Warning: `unite_()` was deprecated in tidyr 1.2.0.
i Please use `unite()` instead.
i The deprecated feature was likely used in the ggmosaic package.
  Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
```



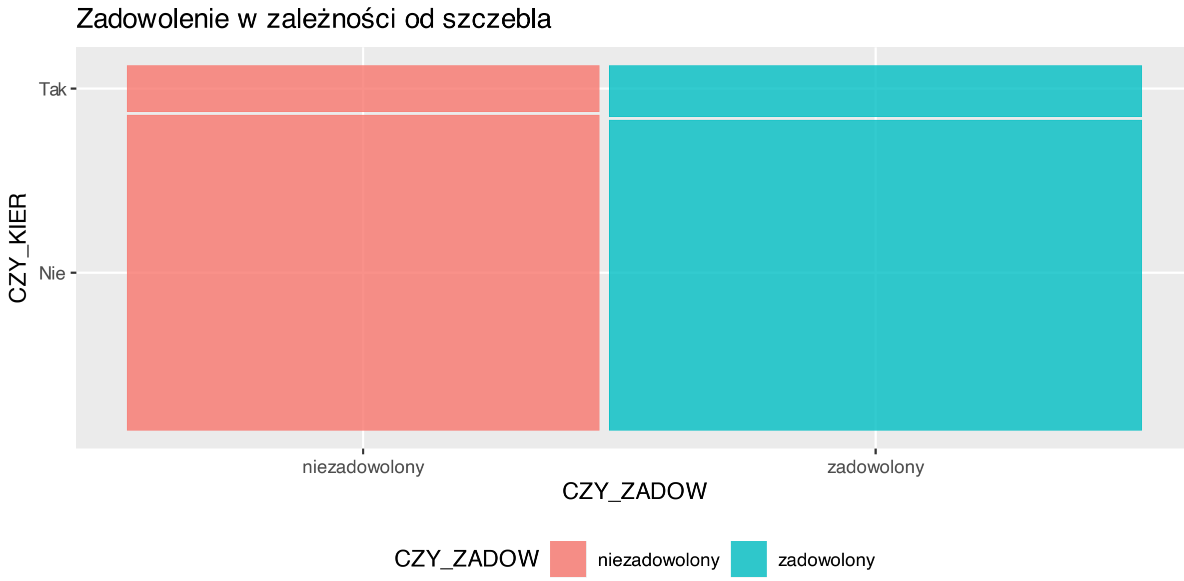
Hipoteza: Dział **HR** ma wpływ na ofertę szkoleń, tym samym może ją dopasować do potrzeb pracowników w tym dziale.

Wykres mozaikowy dla pary zmiennych **CZY_ZADOW** i **STAŻ**.



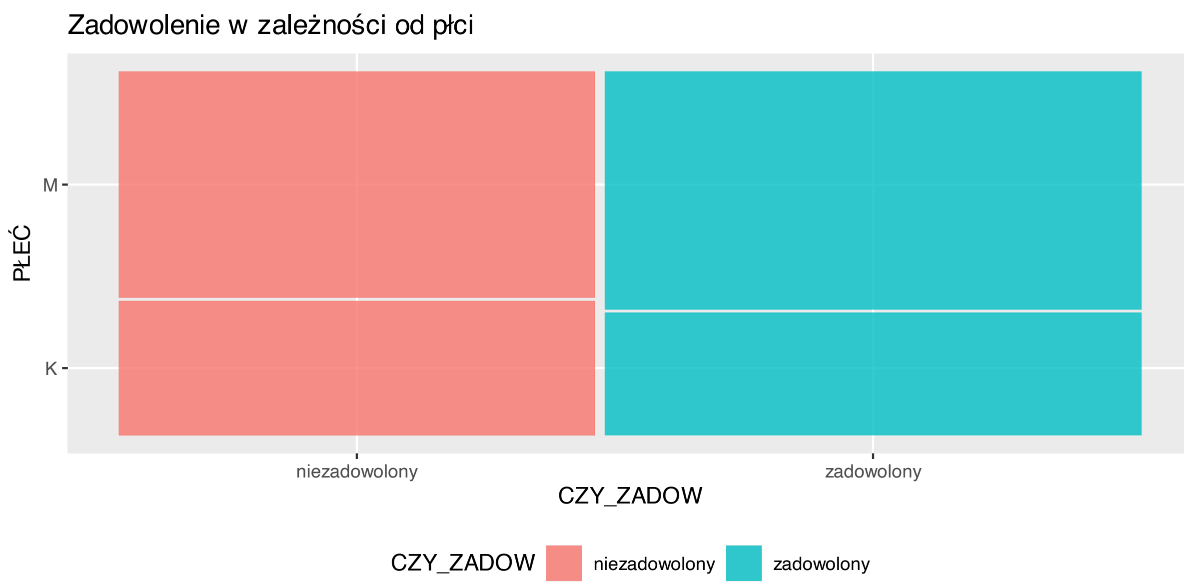
Hipoteza: Osoby zatrudnione w firmie między rokiem a trzema latami mają niedostateczne umiejętności, które mogą być rozwijane w ramach oferowanych szkoleń.

Wykres mozaikowy dla pary zmiennych **CZY_ZADOW** i **CZY_KIER**.



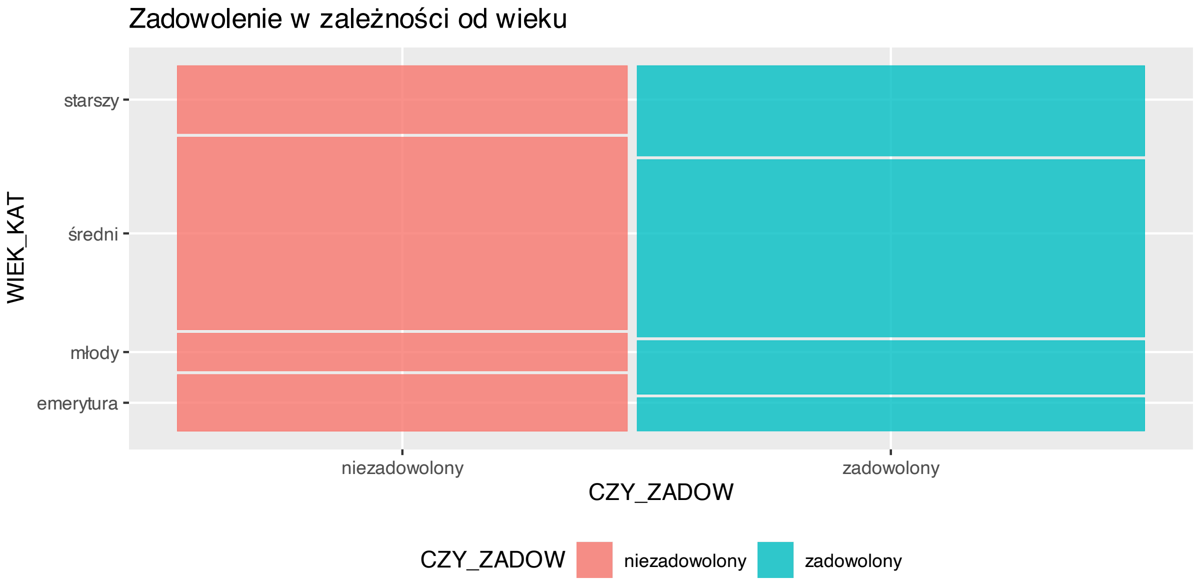
Hipoteza: Zajmowanie przez pracownika stanowiska kierowniczego nie przekłada się na jego opinię o dopasowaniu oferty szkoleń do jego potrzeb zawodowych.

Wykres mozaikowy dla pary zmiennych **CZY_ZADOW** i **PŁEĆ**.



Hipoteza: Kobiety mają wyższe wymagania dotyczące oferty szkoleń w porównaniu do mężczyzn.

Wykres mozaikowy dla pary zmiennych **CZY_ZADOW** i **WIEK_KAT**.



Hipoteza: Oferta szkoleń w firmie jest dobrze dopasowana do osób z wieloletnim doświadczeniem w branży, ale nie uwzględnia w pełni potrzeb osób z doświadczeniem eksperckim.

Część II

Zadanie 2

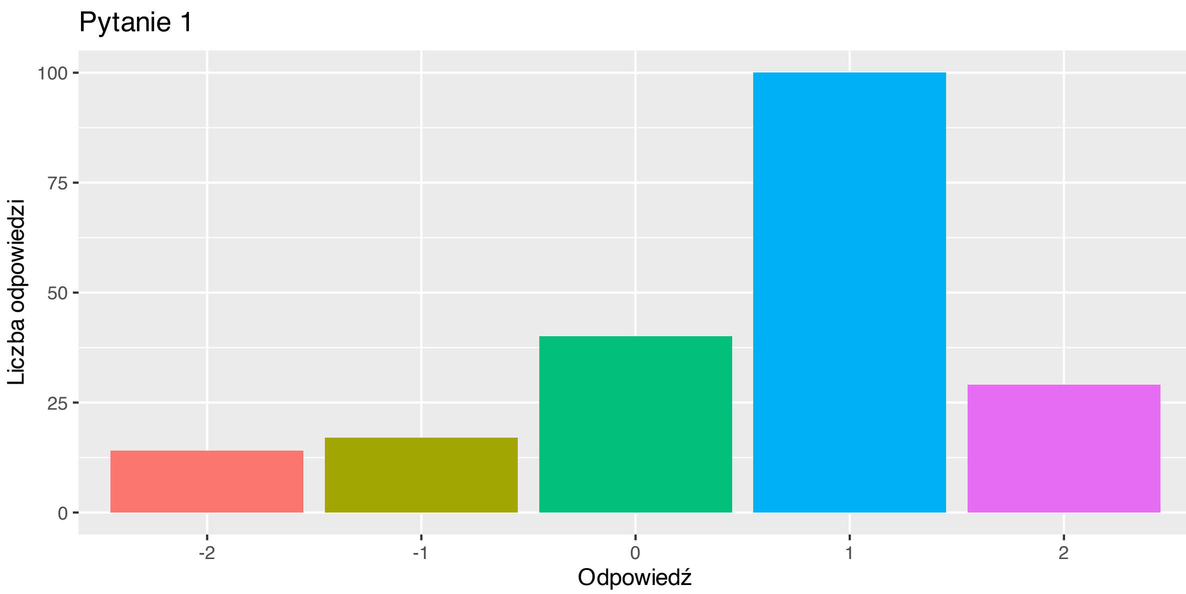
Zilustrowano odpowiedzi na pytanie "Jak bardzo zgadzasz się ze stwierdzeniem, że firma pozwala na (...)?" (zmienna **PYT_1**) w całej badanej grupie oraz w podgrupach ze względu na zmienną **CZY_KIER**.

Wyniki dla całej grupy.

```
ankieta |>
  group_by(PYT_1) |>
  summarise('%' = n() / nrow(ankieta))
```

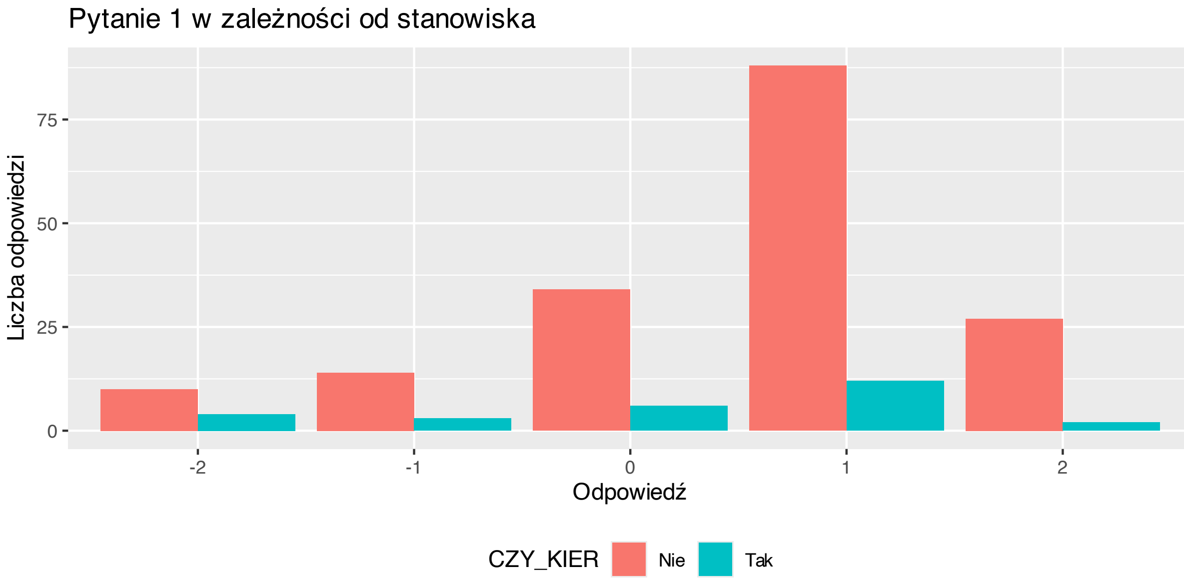
```
# A tibble: 5 x 2
  PYT_1   `%`
  <fct> <dbl>
1 -2    0.07
2 -1    0.085
3 0     0.2
4 1     0.5
5 2     0.145
```

```
ankieta |>
  group_by(PYT_1) |>
  summarise(n=n()) |>
  ggplot(aes(x=PYT_1, y=n, fill=PYT_1)) +
  geom_bar(stat='identity') +
  labs(title='Pytanie 1', x='Odpowiedź', y='Liczba odpowiedzi') +
  theme(legend.position = 'none')
```



Wyniki dla podgrup ze względu na zmienną **CZY_KIER**.

```
ankieta |>
  group_by(PYT_1, CZY_KIER) |>
  summarise(n = n(), .groups='keep') |>
  ggplot(aes(x = PYT_1, y = n, fill = CZY_KIER)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  labs(title='Pytanie 1 w zależności od stanowiska',
       x='Odpowiedź', y='Liczba odpowiedzi') +
  theme(legend.position = 'bottom')
```



Nie widać wyraźniej zależności między zajmowanym stanowiskiem (**CZY_KIER**) a odpowiedzią udzieloną pierwsze pytanie ankiety (**PYT_1**).

Zadanie 3

Zapoznano się z działaniem funkcji *sample* z biblioteki *stats.* Przetestowano jej działanie dla różnych wartości argumentów wejściowych. Następnie wylosowano próbkę o liczności 10% wszystkich rekordów z pliku "ankieta.csv" w dwóch wersjach: ze zwracaniem oraz bez zwracania.

Opisywana funkcja przyjmuje następujące argumenty:

- x – wektor z elementami, z których losujemy, lub liczba naturalna określająca zakres losowania (tj. liczba elementów od 1 do x);
- *size* – liczba elementów do wylosowania;
- *replace* – określa, czy losowanie ma odbywać się ze zwracaniem;
- *prob* – wektor prawdopodobieństw, który przypisuje różne szanse wylosowania poszczególnym elementom;
- *useHash* – wartość logiczna, która określa, czy przy losowaniu ma być stosowana wersja algorytmu wykorzystująca hashowani.

Przykłady działania funkcji *sample* z biblioteki *stats.*

Losowanie ze zwracaniem.

```
example = 1:5
for (n in 2:4) {
  print(sample(example, n, replace=TRUE))
}
```

```
[1] 3 2
[1] 5 1 3
[1] 4 4 5 5
```

Losowanie bez zwracania.

```
for (n in 2:4) {
  print(sample(example, n, replace=FALSE))
}
```

```
[1] 2 4
[1] 3 1 5
[1] 3 5 2 4
```

Próbka wylosowana ze zwracaniem.

```
ankieta[
  sample(nrow(ankieta), nrow(ankieta)*0.1, replace=TRUE),
  ]
```

```
    DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK  WIEK_KAT     CZY_ZADOW
31     PD    1      Nie     1     2     2    M   32     młody     zadowolony
20     IT    2      Nie     1     2     2    K   27     młody     zadowolony
115    PD    1      Nie    -2    -2    -2    M   39    średni niezadowolony
38     PD    1      Nie     0    -2    -1    M   35     młody niezadowolony
82     PD    3      Tak    -1    -2    -2    M   54   starszy niezadowolony
5      IT    3      Tak     1     2    -1    K   65 emerytura     zadowolony
98     PD    2      Nie     0    -2    -2    M   40    średni niezadowolony
167    MK    2      Nie     0    -2    -2    M   58 emerytura niezadowolony
86     PD    2      Tak     1     2     2    M   52   starszy     zadowolony
128    MK    2      Nie     0    -2    -2    K   45    średni niezadowolony
41     PD    1      Nie     1     2     2    M   39    średni     zadowolony
120    PD    2      Nie     1     2     2    M   38    średni     zadowolony
36     PD    1      Nie    -2    -2     1    M   29     młody niezadowolony
28     PD    1      Nie     1     2     2    M   26     młody     zadowolony
```

```
166     MK    2       Tak      1      2      1     M   53   starszy    zadowolony
26      IT    3       Nie      1      2      2     K   51   starszy    zadowolony
141     MK    2       Nie      1      2      2     K   45    średni    zadowolony
59      PD    2       Nie      1     -2     -2     M   49   starszy niezadowolony
178     HR    2       Nie      1     -1      1     M   45    średni niezadowolony
169     MK    2       Nie      2      2      2     M   38    średni    zadowolony
```

Próbka wylosowana bez zwracania.

```r
ankieta[
  sample(nrow(ankieta), nrow(ankieta)*0.1, replace=FALSE),
  ]
```

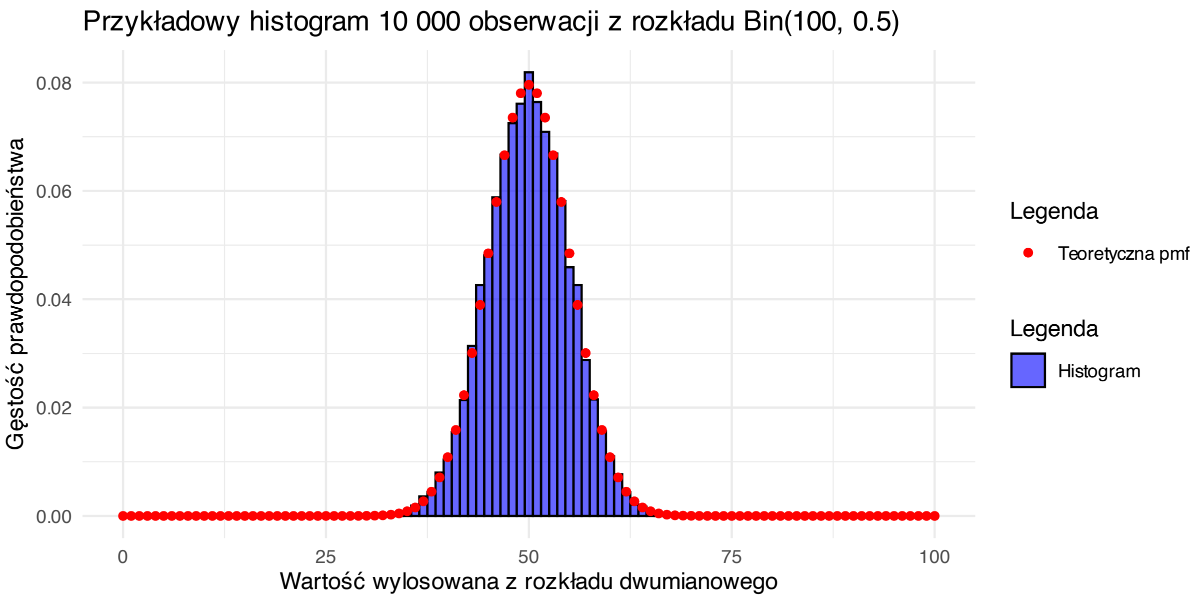
```
    DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK  WIEK_KAT      CZY_ZADOW
13      IT    2       Tak      1      2      2     K   48   starszy    zadowolony
22      IT    3       Nie      1      2     -1     K   64 emerytura    zadowolony
184     HR    2       Nie      1      2      1     M   39    średni    zadowolony
52      PD    1       Nie      2      2      2     M   40    średni    zadowolony
6       IT    3       Tak      0      1      1     K   57 emerytura    zadowolony
142     MK    2       Nie     -1     -2     -1     K   30     młody niezadowolony
104     PD    3       Tak      1      2      2     K   36    średni    zadowolony
153     MK    2       Nie      1     -1     -1     M   65 emerytura niezadowolony
179     HR    2       Nie      0     -2     -1     M   38    średni niezadowolony
183     HR    2       Nie      2      2      2     M   38    średni    zadowolony
175     HR    2       Nie      1      2      2     M   43    średni    zadowolony
189     HR    2       Nie      1      2      2     M   49   starszy    zadowolony
132     MK    3       Tak      0     -2     -2     K   42    średni niezadowolony
100     PD    2       Nie      1      2      1     K   41    średni    zadowolony
137     MK    2       Nie      2      2      2     K   41    średni    zadowolony
60      PD    2       Nie      1     -1      1     M   55   starszy niezadowolony
157     MK    2       Nie      1      2      2     M   49   starszy    zadowolony
84      PD    1       Nie      1      2      2     M   54   starszy    zadowolony
148     MK    2       Nie      1      2      2     K   48   starszy    zadowolony
150     MK    3       Tak      0     -2     -2     M   37    średni niezadowolony
```

Zadanie 4

Zaproponowano metodę symulowania zmiennych losowych z rozkładu dwumianowego. Napisano funkcję do generowania realizacji, a następnie zaprezentowano jej działanie porównując wybrane teoretyczne i empiryczne charakterystyki dla przykładowych wartości parametrów rozkładu: n i p.

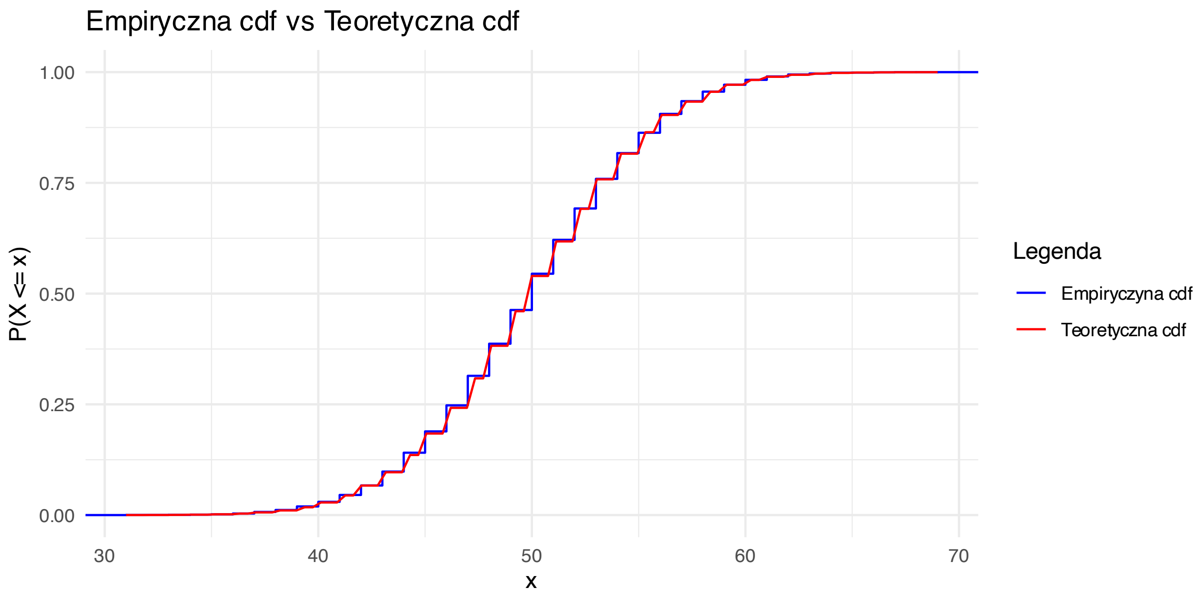
```r
bin_rvs <- function(n, p) {
  sum(sample(c(0, 1), n, replace=TRUE, prob=c(1-p, p)))
 }
```

```r
n <- 100
p <- 0.5
xs <- tibble(Value=replicate(100*n, bin_rvs(n, p)))
ggplot(xs, aes(x=Value)) +
  geom_histogram(
    aes(y=after_stat(density), fill="Histogram"),
    bins=n+1,
    color="black",
    binwidth = 1,
    alpha=0.6
    ) +
  stat_function(
    fun=dbinom,
    aes(color='Teoretyczna pmf', fill='Teoretyczna pmf'),
    xlim=c(0, 100),
    args=list(size=n, prob=p),
    geom='point',
    n=101
    ) +
  ggtitle(
    "Przykładowy histogram 10 000 obserwacji z rozkładu Bin(100, 0.5)"
    ) +
  ylab("Gęstość prawdopodobieństwa") +
  xlab("Wartość wylosowana z rozkładu dwumianowego") +
   scale_fill_manual(
    name = "Legenda",
    values = c("Histogram" = "blue")
  ) +
  scale_color_manual(
    name = "Legenda",
    values = c('Teoretyczna pmf' = "red")
  ) +
  theme_minimal()
```



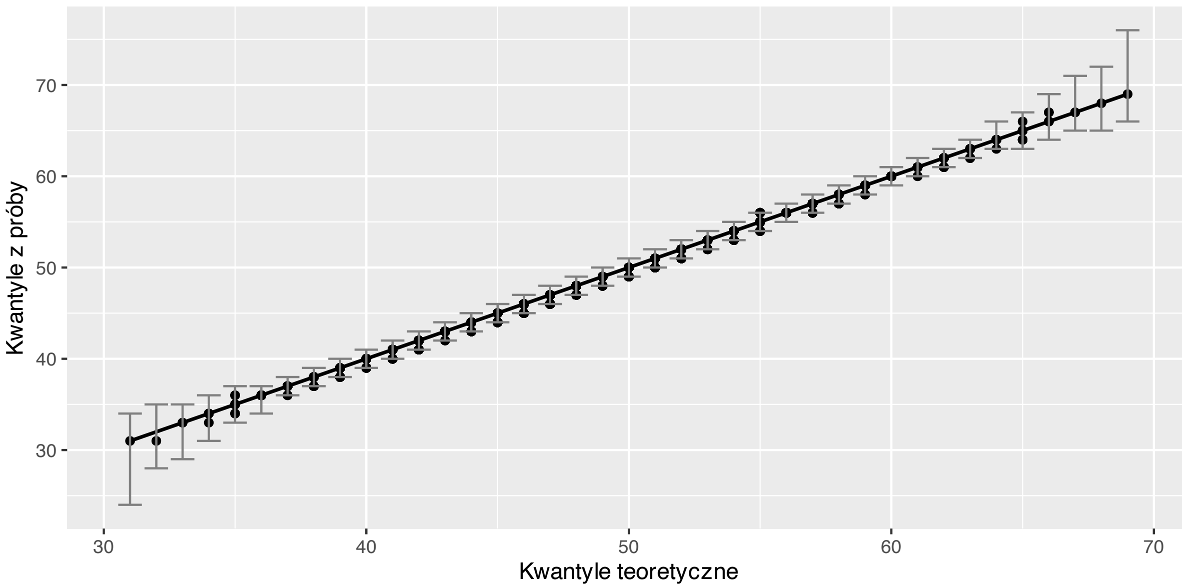
Histogram dla wysymulowanej próbki oraz teoretyczne wartości funkcji masy prawdopodobieństwa (pmf) z rozkładu dwumianowego o parametrach n=100 i p=0.5 pokrywają się ze sobą.

```
ggplot(xs, aes(x=Value)) +
  stat_ecdf(aes(color = 'Empiryczyna cdf'), geom="step") +
  stat_function(
    fun=pbinom,
    aes(color = 'Teoretyczna cdf'),
    args=list(size=n, prob=p)
    ) +
  ggtitle("Empiryczna cdf vs Teoretyczna cdf") +
  ylab("P(X <= x)") +
  xlab("x") +
  scale_color_manual(
    name = "Legenda",
    values = c('Empiryczyna cdf' = "blue", 'Teoretyczna cdf' = "red")
  ) +
  theme_minimal()
```



Empiryczna oraz teoretyczna dystrybuanta dla rozkładu dwumianowego o parametrach n=100 i p=0.5 pokrywają się ze sobą.

```
ggplot(xs, aes(sample=Value)) +
  stat_qq_point(
    distribution="binom", dparams=list(size=n, prob=p)
    ) +
  stat_qq_line(
    distribution="binom", dparams=list(size=n, prob=p)
    ) +
  stat_qq_band(
    distribution="binom", dparams=list(size=n, prob=p), bandType="ell"
    ) +
  ylab("Kwantyle z próby") +
  xlab("Kwantyle teoretyczne")
```



Wyznaczone punkty na wykresie mieszą się w przedziałach ufności wyznaczonych przez funkcję *stat_qq_band*. Zatem na poziomie istotności 0.05 nie mamy podstaw do odrzucenia hipotezy o zgodności rozkładów.

```
# teoretyczna wartość oczekiwana
n * p
```

```
[1] 50
```

```
# empiryczna wartość oczekiwana
mean(xs$Value)
```

```
[1] 49.9462
```

```
# teoretyczna wariancja
n * p * (1 - p)
```

```
[1] 25
```

```
# empiryczna wariancja
var(xs)
```

```
         Value
Value 25.23003
```

Porównanie teoretycznych i empirycznych charakterystyk dla rozkładu *Bin(100; 0.5)* wskazuje, że zaimplementowana funkcja działa poprawnie.

Zadanie 5

Zaproponowano metodę symulowania wektorów losowych z rozkładu wielomianowego. Napisano funkcję do generowania realizacji, a następnie zaprezentowano jej działanie porównując teoretyczne prawdopodobieństwa z wyliczonymi empirycznie proporcjami dla przykładowych wartości paramertrów rozkładu: n *i* p.

```r
multinomial_rv <- function(n, p) {
  X <- rep(0, length(p))
  for (i in 1:n) {
    temp <- sample(1:length(p), 1, prob=p)
    X[temp] <- X[temp] + 1
  }
  X
}
```

```r
multinomial_rvs <- function(size, n, p) {
  matrix(
    replicate(size, multinomial_rv(n, p)), nrow=length(p)
    )
}
```

```r
p <- c(0.1, 0.2, 0.3, 0.4)
n <- 1000
size <- 1
x <- multinomial_rvs(size, n, p)
rowSums(x) / size / n  # empiryczne prawdopodbieństwa
```

```
[1] 0.093 0.200 0.292 0.415
```

```r
p  # teoretyczne prawdopodobieństwa
```

```
[1] 0.1 0.2 0.3 0.4
```

Porównanie teoretycznych prawdopodobieństw z empirycznymi proporcjami, obliczonymi na podstawie jednokrotnego losowania 1000 elementów zgodnie z zadanym wektorem prawdopodobieństw (0.1, 0.2, 0.3, 0.4), wskazuje, że zaimplementowana funkcja działa poprawnie.

Część III oraz IV

Zadanie 6

Napisano funkcję do wyznaczania realizacji przedziału ufności Cloppera–Pearsona. Argumentem wejściowym jest poziom ufności, liczba sukcesów i liczba prób lub poziom ufności i wektor danych (funkcja obsługuje oba przypadki).

```
CP_CI <- function(conf.level, x, n=NA) {
  if (is.na(n)) {
    dane <- x
    x <- sum(dane=='zadowolony')
    n <- length(dane)
  }
  alpha <- 1 - conf.level
  L <- qbeta(alpha / 2, x, n - x + 1)
  U <- qbeta(1 - alpha / 2, x + 1, n - x)
  if (x == 0) {
    L <- 0
  }
  if (x == n) {
    U <- 1
  }
  data.frame(est=x/n, lwr.ci=L, upr.ci=U)
}
```

Zadanie 7

Korzystając z funkcji napisanej w zadaniu 6. wyznaczono realizacje przedziałów ufności dla prawdopodobieństwa, że pracownik uważa szkolenia za przystosowane do swoich potrzeb w pierwszym badanym okresie oraz w drugim badanym okresie. Skorzystano ze zmiennych **CZY_ZADW** oraz **CZY_ZADW_2** (zmienną utworzono analogicznie jak w zadaniu 1.7). Przyjęto $1 - \alpha = 0.95$.

```
ankieta['CZY_ZADOW_2'] <- ifelse(ankieta$PYT_3 %in% c(1, 2),
                                  'zadowolony', 'niezadowolony')
x_zadw <- ankieta |>
  filter(CZY_ZADOW == 'zadowolony') |> nrow()
x_zadw2 <- ankieta |>
  filter(CZY_ZADOW_2 == 'zadowolony') |> nrow()
n <- nrow(ankieta)
```

Poniżej przedstawiono estymowane wartości prawdopodobieństw, realizacje dolnej oraz górnej granicy przedziału ufności otrzymane dla obu badanych okresów.

Estymowana wartość prawdopodobieństwa zadowolenia oraz realizacja przedziału ufności dla prawdopodobieństwa zadowolenia w pierwszym okresie wyznaczona przy pomocy zaimplementowanej funkcji.

```
CP_CI(0.95, ankieta$CZY_ZADOW)
```

```
   est    lwr.ci    upr.ci
1 0.53 0.4583305 0.6007671
```

Estymowana wartość prawdopodobieństwa zadowolenia oraz realizacja przedziału ufności dla prawdopodobieństwa zadowolenia w pierwszym okresie wyznaczona przy pomocy funkcji bibliotecznej.

```
BinomCI(
  x_zadw, n, method="clopper-pearson", conf.level=0.95
  )
```

```
      est    lwr.ci    upr.ci
[1,] 0.53 0.4583305 0.6007671
```

Estymowana wartość prawdopodobieństwa zadowolenia oraz realizacja przedziału ufności dla prawdopodobieństwa zadowolenia w drugim okresie wyznaczona przy pomocy zaimplementowanej funkcji.

```
CP_CI(0.95, x_zadw2, n)
```

```
   est    lwr.ci    upr.ci
1 0.59 0.5184216 0.6588694
```

Estymowana wartość prawdopodobieństwa zadowolenia oraz realizacja przedziału ufności dla prawdopodobieństwa zadowolenia w drugim okresie wyznaczona przy pomocy funkcji bibliotecznej.

```
BinomCI(
  x_zadw2, n, method="clopper-pearson", conf.level=0.95
  )
```

```
      est    lwr.ci    upr.ci
[1,] 0.59 0.5184216 0.6588694
```

Wyniki dla zaimplementowanej oraz bibliotecznej funkcji były sobie równe dla obu rozważanych terminów. Oznacza to, że funkcja napisana w zadaniu 6. działa poprawnie.

Zadanie 8

Zapoznano się z funkcjami do generowania zmiennych losowych z rozkładu dwumianowego oraz do wyznaczania przedziałów ufności dla parametru p. Przetestowano ich działanie.

W pakiecie R do wygenerowania zmiennych losowych z rozkładu dwumianowego można posłużyć się funkcją *rbinom*, znajdującą się w bibliotece *stats*.

Funkcja jako argumenty przyjmuje:

- n – liczba losowanych wartości,
- *size* – liczba prób w pojedynczej realizacji,
- *prob* – prawdopodobieństwo sukcesu w każdej próbie.

Poniżej znajduje się przykład użycia omawianej funkcji. Wylosowano 10 wartości z rozkładu *Bin(100; 0.5)*.

```r
rbinom(10, 100, 0.5)
```

```
 [1] 55 55 45 46 61 47 54 50 49 61
```

Do wyznaczenia przedziałów ufności dla parametru p służy funkcja *BinomCl* z biblioteki *DescTools*.

Funkcja jako argumenty przyjmuje:

- x – liczba sukcesów;
- n – liczba prób;
- *conf.level* – poziom istotności, domyślnie 0.95;
- *sides* – strona przedziału ufności, domyślnie “two.sided”;
- *methods* – metoda obliczania przedziału ufności, domyślnie “wilson”.

Poniżej znajduje się przykład użycia omawianej funkcji. Wynikiem jest realizacja obustronnego przedziału ufności na poziomie ufności 0.95 wyznaczonego metodą Cloppera – Pearsona przy liczbie sukcesów wynoszącej 50 i liczbie prób równej 100.

```
BinomCI(50, 100, method="clopper-pearson")
```

```
     est    lwr.ci    upr.ci
[1,] 0.5 0.3983211 0.6016789
```

Zadanie 9

Przeprowadzono symulacje, których celem jest porównanie prawdopodobieństwa pokrycia i długości przedziałów ufności Cloppera-Pearsona, Walda i Jeffreysa. Rozważono 1 - $\alpha = 0.95$, rozmiar próby n $\in \{30,100,1000\}$ i różne wartości prawdopodobieństwa p. Wyniki umieszczono na wykresach i sformułowano wnioski, które dla konkretnych danych ułatwiają wybór konkretnego typu przedziału ufności.

```
methods <- c("clopper-pearson", "wald", "jeffreys")
ns <- c(30, 100, 1000)
ps <- seq(from = 0.01, to = 0.99, by = 0.01)
```

Funkcja *sprawdzaj_CI_exact* dla danej metody zwraca listę zawierającą trzy elementy :

- *valid* – zmienna określająca, czy prawdopodobieństwo pokrycia jest równe co najmniej 0.95,
- *length* – oczekiwaną długość przedziału ufności,
- *coverage* – prawdopodobieństwo pokrycia.

Funkcja liczy wartość oczekiwaną długości przedziału ufności sumując długości przedziałów ufności dla wszystkich możliwych wartości parametru p ważonych prawdopodobieństwem wystąpienia danej wartości dla ustalonych n i p. Następnie oblicza prawdopodbieństwo pokrycia, sumując prawdopodobieństwa, dla których dany przedział ufności zawiera daną wartość p.

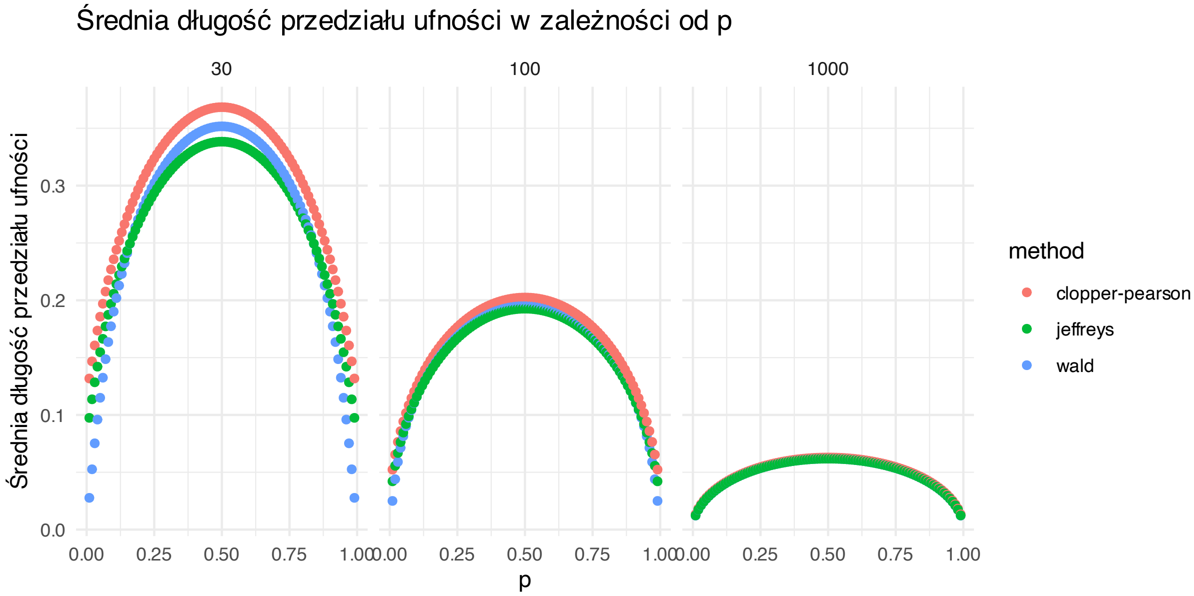
```
sprawdzaj_CI_exact <- function(n, p, method, alpha=0.05){
  X <- dbinom(0:n, n, p)
  wyniki <- BinomCI(0:n, n, method=method, conf.level=1-alpha)
  dlugosc <- sum((wyniki[,'upr.ci'] - wyniki[,'lwr.ci']) * X)
  Y <- sum(1*((wyniki[, 'lwr.ci'] <= p) & (p <= wyniki[, 'upr.ci'])) * X)
  list(valid=Y>=1-alpha, length=dlugosc, coverage=Y)
}
```

```
results <- data.frame(method=c(), n=c(), p=c(), valid=c(),
                      length=c(), coverage=c())
for (n in ns){
  for (p in ps){
    for (method in methods){
      res <- sprawdzaj_CI_exact(n, p, method)
      results <- rbind(results, data.frame(method=method, n=n, p=p, res))
  }}}
```

Przy interpretacji wyników dla danej metody interesować nas będzie zarówno średnia długość przedziału, jak i prawdopodobieństwo pokrycia dla parametru p. Poniżej zaprezentowano trzy wykresy: pierwszy z nich przedstawia średnią długość przedziału w zależności od wartości p, drugi wykres pokazuje poziom pokrycia w zależności od p. Na końcu natomiast znajduje się tzw. "studnia" – dla ustalonej paru parametrów n i p wskazuje ona metodę, dla której średnią długość jest najmniejsza oraz prawdopodopodobieństwo pokrycia jest równe co najmniej zadanemu poziomowi ufności.

Przy wykorzystaniu funkcji *sprawdzaj_CI_exact*, naniesiono na wykres średnią długość przedziału ufności w zależności od p dla wszystkich badanych metod oraz wartości n.

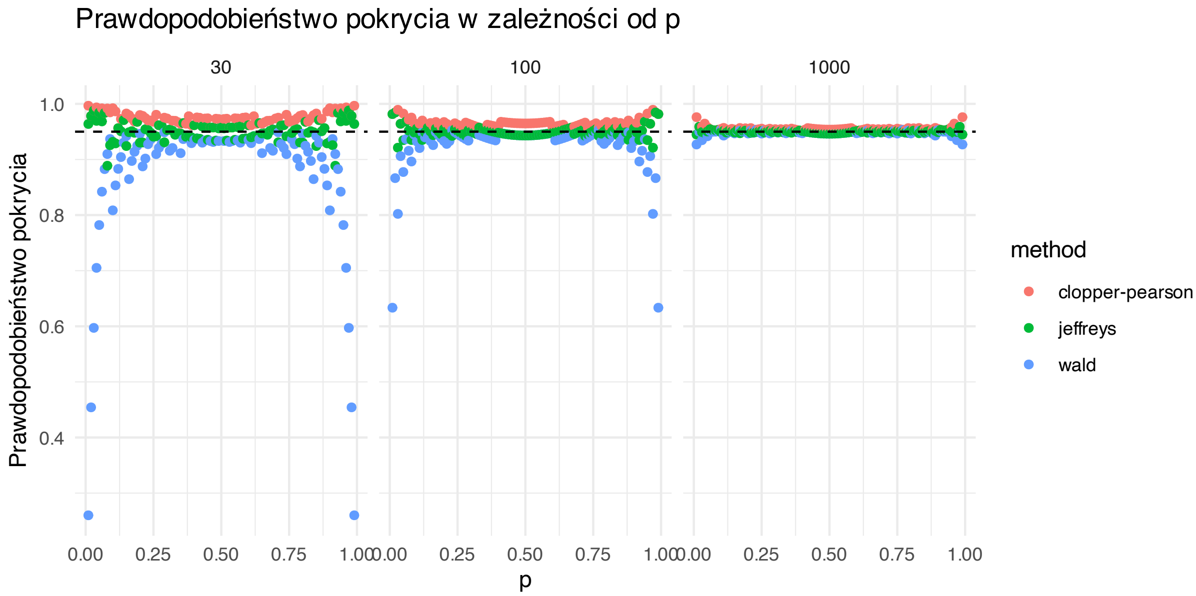
```
results |>
  ggplot(aes(x=p, y=length, color=method)) +
  geom_point() +
  labs(title='Średnia długość przedziału ufności w zależności od p',
       x='p', y='Średnia długość przedziału ufności') +
  theme_minimal() +
  facet_wrap(~n)
```



Najkrótszym przedziałem ufności najczęściej okazuje się przedział Jeffreysa. Nieco gorszy jest przedział Walda, natomiast przedział Cloppera-Pearsona wypada najgorzej. Różnice widoczne są dla małej wielkości próby (n=30) i zaczynają zanikać wraz z jej wzrostem.

Przy wykorzystaniu funkcji *sprawdzaj_CI_exact*, naniesiono na wykres prawdopodobieństwo pokrycia w zależności od p dla wszytskich badanych metod oraz wartości n.

```
results |>
  ggplot(aes(x=p, y=coverage, color=method)) +
  geom_point() +
  labs(title='Prawdopodobieństwo pokrycia w zależności od p',
       x='p', y='Prawdopodobieństwo pokrycia') +
  theme_minimal() +
  facet_wrap(~n) +
  geom_hline(yintercept=0.95, linetype='dashed')
```

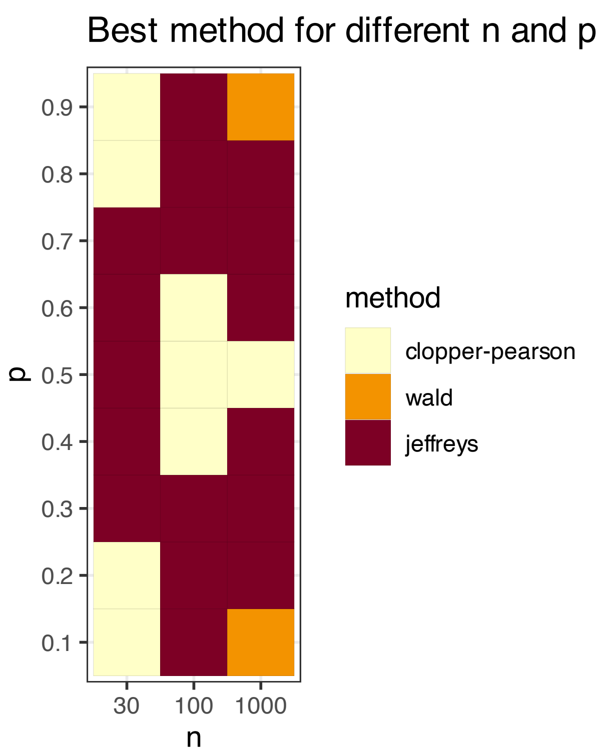


Prawdopodobieństwo pokrycia dla przedziału opartego na metodzie Cloppera-Pearsona dla każdej pary parametrów jest równa co najmniej zadanemu poziomowi ufności. Przedział Jeffreysa dla większości przypadków także zapewniał odpowiednie pokrycie. Przedział Walda wypadł w tym zestawieniu najgorzej. Podobnie jak w przypadku wykresu dotyczącego średniej długości przedziału, różnice te zaczynają się zacierać wraz ze wzrostem rozmiaru próby.

Przy wykorzystaniu funkcji *sprawdzaj_CI_exact* stworzono tzw. "studnię" – dla każdej pary parametrów sprawdzono, czy dana metoda zapewnia odpowiedni poziom pokrycia, a następnie, spośród kwalifikujących się metod, wybrano tę o najkrótszej średniej długości przedziału.

```
ps <- c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)

best_of_methods <- function(n, p, methods, ...){
  ramka <- data.frame(valid=c(), length=c(), coverage=c())
  for (method in methods) {
    ramka <- tryCatch({
      rbind(ramka, as.data.frame(sprawdzaj_CI_exact(n, p, method, ...)))
    }, error=function(...){
      rbind(ramka, data.frame(valid=FALSE, length=NA, coverage=NA))
    })
  }
  ramka$method <- methods
  valid_methods <- ramka[ramka$valid, ]
  which(valid_methods[which.min(valid_methods$length),]$method == methods)
}
```

```r
wyniki <- matrix(NA, nrow=length(ns), ncol=length(ps))
for (i in 1:length(ns)){
  for (j in 1:length(ps)){
    wyniki[i, j] <- best_of_methods(ns[i], ps[j], methods)
  }
}
rownames(wyniki) <- ns
colnames(wyniki) <- ps
data <- expand.grid(Var1=rownames(wyniki), Var2=colnames(wyniki))
data$value <- as.vector(wyniki)
data$method <- methods[data$value]
colors = hcl.colors(length(methods), 'ylorRd', rev=TRUE)
ggplot(data, aes(x=Var1, y=Var2, fill=factor(value))) +
  geom_tile() +
  geom_tile(color = "#00000022") +
  scale_fill_manual(values = colors, labels=methods)+
  coord_equal() +
  theme_bw() +
  labs(title="Best method for different n and p",
       x="n", y="p", fill="method")
```



Wykres należy interpretować, sprawdzając kolor odpowiedniego kwadratu i odczytując przypisaną mu metodę z legendy.

W większości przypadków najlepszym wyborem okazał się przedział oparty na metodzie Jeffreysa.

Drugie miejsce zajęła metoda Cloppera–Pearsona, która sprawdziła się przy małej wielkości próby (n=30) oraz skrajnych wartości parametru p (bliskich 0 lub 1). Była również skuteczna przy większych próbach ($n \in \{100, 1000\}$), gdy p było bliskie 0.5.

Przedział otrzymany za pomocą metody Walda był najlepszy tylko dla dużej próby (n=1000) i skrajnych wartości prawdopodobieństwa ($p \in \{0.1, 0.9\}$).

Część V

Zadanie 10

Zapoznano się z funkcjami służącymi do wykonania testu dokładnego oraz asympotycznego weryfikujcego hipotezę zerową dotyczcą prawodopodobiestwa sukcesu z rozkładu dwumianowego.

Do wykonania testu dokładnego weryfikujcego hipotezę zerową dotyczcą prawodopodobieństwa sukcesu z rozkładu dwumianowego można posłużyć się funkcją *binom.test* z biblioteki *stats*. Za argumenty przyjmuje ona:

- x – liczba sukcesów lub wektor dwuelementowy zawierający kolejno liczbę sukcesów i porażek;
- n – liczba prób, ignorowane jeśli x jest wektorem o długości 2;
- p – zakładane prawdopodobieństwo sukcesu;
- *alternative* – określa hipotezę alternatywną, dostępne opcje to: “two.sided”, “greater”, “less”;
- *conf.level* – poziom ufności dla zwróconego przedziału ufności.

Funkcja zwraca *p–value*, pozwalającą ocenić, czy są podstawy do odrzucenia hipotezy zerowej. W wyniku znajdziemy także estymowane prawdopodobieństwo sukcesu oraz realizację przedziału ufności dla $\widehat{p}$ na poziomie ufności *conf.level*.

Do wykonania testu asymptotycznego weryfikujcego hipotezę zerową dotyczcą prawodopodobiestwa sukcesu z rozkładu dwumianowego można posłużyć się funkcją *prop.test* z biblioteki *stats*. Za jej pomocą możemy zbadać czy prawdopodobieństwa sukcesów wśród kilku grup są sobie równe, bądź czy równają się zadanym wartościom. Za argumenty przyjmuje ona:

- x – wektor z liczbą sukcesów lub macierz z dwoma kolumnami odpowiadającymi kolejno liczbie sukcesów i porażek;
- n – wektor z liczba prób, ignorowane jeśli x jest macierzą;
- p – zakładane prawdopodobieństwa sukcesów;
- *alternative* – określa hipotezę alternatywną, dostępne opcje to: "two.sided", "greater", "less". Nieignorowane tylko w przypadku, gdy weryfikujemy hipotezę dla jednej grupy lub sprawdzamy, czy prawdopodobieństwa sukcesów dla dwóch grup są sobie równe;
- *conf.level* – poziom ufności dla zwróconego przedziału ufności.Nieignorowane tylko w przypadku, gdy weryfikujemy hipotezę dla jednej grupy lub sprawdzamy, czy prawdopodobieństwa sukcesów dla dwóch grup są sobie równe;
- *correct* – wartość logiczna wskazująca, czy korekta ciągłości Yatesa powinna być stosowana tam, gdzie to możliwe.

Przy testowaniu hipotez funkcja zwraca wartość statystyki testowej *X-squared*, *p–value*, a także estymowane prawdopodobieństwa sukcesu. Przy badaniu jednej grupy otrzymamy oraz realizację przedziału ufności dla $\widehat{p}$ na poziomie ufności *conf.level*. W przypadku dwóch grup przedział dotyczy różnicy prawdopodobieństw sukcesów.

Zadanie 11

Dla danych z pliku "ankieta.csv" korzystając z funkcji z zadania 10., przyjmując $1-\alpha = 0.95$, zweryfikowano następujące hipotezy:

1. Prawdopodobieństwo, że w firmie pracuje kobieta wynosi 0.5.

```
exact_results <- binom.test(
  length(ankieta$PŁEĆ[ankieta$PŁEĆ == 'K']),
  nrow(ankieta),
  p=0.5,
  conf.level=conf.level
  )
asymptotic_results <-prop.test(
  length(ankieta$PŁEĆ[ankieta$PŁEĆ == 'K']),
  nrow(ankieta),
  p=0.5,
  conf.level=conf.level
  )
```

Wyniki testu dokładnego

```
    Exact binomial test

data:  length(ankieta$PŁEĆ[ankieta$PŁEĆ == "K"]) and nrow(ankieta)
number of successes = 71, number of trials = 200, p-value = 4.973e-05
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2887838 0.4255862
sample estimates:
probability of success
                 0.355
```

Wyniki testu asymptotycznego

```
    1-sample proportions test with continuity correction

data:  length(ankieta$PŁEĆ[ankieta$PŁEĆ == "K"]) out of nrow(ankieta), null probability 0.5
X-squared = 16.245, df = 1, p-value = 5.566e-05
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2896363 0.4260327
sample estimates:
    p
0.355
```

Wyniku obu testów ($p - value < 0.05$) wskazują na to, że na poziomie istotności $\alpha = 5\%$ należy odrzucić hipotezę zerową na rzecz hipotezy alternatywnej. Przyjmujemy, że prawdopodobieństwo, że w firmie pracuje kobieta nie wynosi 0.5. Estymowana wartość prawdopodobieństwa wynosi 0.355.

2. Prawdopodbieństwo, że pracownik uważa szkolenia za przystosowane do swoich potrzeb w pierwszym badanym okresie jest większe bądź równe 0.7.

```r
exact_results <- binom.test(
  length(ankieta$PYT_2[ankieta$PYT_2 %in% c(1, 2)]),
  nrow(ankieta),
  p=0.7,
  conf.level=conf.level,
  alternative='less')

asymptotic_results <- prop.test(
```

```r
  length(ankieta$PYT_2[ankieta$PYT_2 %in% c(1, 2)]),
  nrow(ankieta),
  p=0.7,
  conf.level=conf.level,
  alternative='less')
```

Wyniki testu dokładnego

```
    Exact binomial test

data:  length(ankieta$PYT_2[ankieta$PYT_2 %in% c(1, 2)]) and nrow(ankieta)
number of successes = 106, number of trials = 200, p-value = 3.213e-07
alternative hypothesis: true probability of success is less than 0.7
95 percent confidence interval:
 0.0000000 0.5899194
sample estimates:
probability of success
                  0.53
```

Wyniki testu asymptotycznego

```
    1-sample proportions test with continuity correction

data:  length(ankieta$PYT_2[ankieta$PYT_2 %in% c(1, 2)]) out of nrow(ankieta), null probabil
X-squared = 26.72, df = 1, p-value = 1.176e-07
alternative hypothesis: true p is less than 0.7
95 percent confidence interval:
 0.0000000 0.5897106
sample estimates:
   p
0.53
```

Wyniku obu testów ($p - value < 0.05$) wskazują na to, że na poziomie istotności $\alpha = 5\%$ należy odrzucić hipotezę zerową na rzecz hipotezy alternatywnej. Prawdopodobieństwo, że pracownik uważa szkolenia za przystosowane do swoich potrzeb w pierwszym badanym okresie jest mniejsze niż 0.7. Estymowana wartość prawdopodobieństwa wynosi 0.53.

3. Prawdopodobieństwo, że kobieta pracuje na stanowisku kierowniczym jest równe prawdopodobieństwu, że mężczyzna pracuje na stanowisku kierowniczym.

```
x1 <- subset(ankieta, CZY_KIER == 'Tak' & PŁEĆ == 'K') |> nrow()
x2 <- subset(ankieta, CZY_KIER == 'Tak' & PŁEĆ == 'M') |> nrow()
n1 <- subset(ankieta, PŁEĆ == 'K' ) |> nrow()
n2 <- subset(ankieta, PŁEĆ == 'M' ) |> nrow()
results <- prop.test(
  c(x1, x2),
  c(n1, n2),
  conf.level=conf.level,
  alternative='two.sided'
  )
```

Wyniki testu asymptotycznego

```
    2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 0.22014, df = 1, p-value = 0.6389
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1411817  0.0719602
sample estimates:
   prop 1    prop 2
0.1126761 0.1472868
```

Wynik testu wskazuje ($p-value > 0.05$), że na poziomie istotności $\alpha = 5\%$ nie mamy podstaw do odrzucenia hipotezy zerowej o równości prawdopodobieństw na rzecz hipotezy alternatywnej.

4. Prawdopodobieństwo, że kobieta uważa szkolenia za przystosowane do swoich potrzeb w pierwszym badanym okresie jest równe prawdopodobieństwu, że mężczyzna uważa szkolenia za przystosowane do swoich potrzeb w pierwszym badanym okresie.

```
x1 <- subset(ankieta, PYT_1 %in% c(1, 2) & PŁEĆ == 'K') |> nrow()
x2 <- subset(ankieta, PYT_1 %in% c(1, 2) & PŁEĆ == 'M') |> nrow()
n1 <- subset(ankieta, PŁEĆ == 'K') |> nrow()
n2 <- subset(ankieta, PŁEĆ == 'M') |> nrow()
results <- prop.test(
  c(x1, x2),
  c(n1, n2),
  conf.level=conf.level,
```

```
  alternative='two.sided'
  )
```

Wyniki testu asymptotycznego

```
    2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 0.047399, df = 1, p-value = 0.8277
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1224584  0.1750843
sample estimates:
   prop 1    prop 2
0.6619718 0.6356589
```

Wynik testu wskazuje ($p-value > 0.05$), że na poziomie istotności $\alpha = 5\%$ nie mamy podstaw do odrzucenia hipotezy zerowej o równości prawdopodobieństw na rzecz hipotezy alternatywnej.

5. Prawdopodobieństwo, że kobieta pracuje w dziale zasobów ludzkich jest większe lub równe prawdopodobieństwu, że mężczyzna pracuje w dziale zasobów ludzkich.

```
x1 <- subset(ankieta, DZIAŁ == 'HR' & PŁEĆ == 'K') |> nrow()
x2 <- subset(ankieta, DZIAŁ == 'HR' & PŁEĆ == 'M') |> nrow()
n <- subset(ankieta, DZIAŁ == 'HR') |> nrow()
results <- prop.test(
  c(x1, x2),
  c(n, n),
  conf.level=conf.level,
  alternative='less'
  )
```

Wyniki testu asymptotycznego

```
    2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n, n)
X-squared = 31.226, df = 1, p-value = 1.148e-08
```

```
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000 -0.5696182
sample estimates:
   prop 1    prop 2
0.1290323 0.8709677
```

Wynik testu wskazuje ($p-value < 0.05$), że na poziomie istotności $\alpha = 5\%$ powinniśmy odrzucić hipotezę zerową na rzecz hipotezy alternatywnej. Odrzucenie hipotezy zerowej sugeruje, że prawdopodobieństwo, iż kobieta pracuje w dziale zasobów ludzkich, jest istotnie mniejsze niż prawdopodobieństwo, że pracuje tam mężczyzna.

Zadanie 12

Wyznaczono symulacyjnie moc testu dokładnego oraz moc testu asymptotycznego w przypadku weryfikacji hipotezy zerowej $H_0 : p = 0.9$ przeciwko $H_1 : p \neq 0.9$ przyjmując wartość $1 - \alpha = 0.95$. Uwzględniono różne wartości alternatyw i różne rozmiary próby.

Zdefiniowana została funkcja *test*, która przyjmuje liczbę sukcesów i rozmiar próby, a następnie wykonuje dwa testy: dokładny i asymptotyczny. Dla każdego z nich porównuje wartość *p-value* z poziomem istotności i zwraca decyzję o odrzuceniu lub nieodrzuceniu hipotezy.

```r
p0 <- 0.9
alpha <- 0.05

test <- function(x, n){
  p_hat <- x/n
  H_exact <- ifelse(
    binom.test(x, n, p=p0)$p.value < alpha, 'H1', 'H0'
    )
  H_asymptotic <- ifelse(
    prop.test(x, n, p=p0)$p.value < alpha, 'H1', 'H0'
    )
data.frame(method = c('H_exact', 'H_asymptotic'),
           H = c(H_exact, H_asymptotic))
}
```

Funkcja *power* służy do oszacowania mocy testu poprzez symulację Monte Carlo. Dla ustalonego rzeczywistego poziomu p_{value} i rozmiaru próby n, generowane jest $N = 100$ prób losowych, a następnie sprawdzane jest, jak często każdy z testów odrzuca hipotezę zerową. Proporcja tych przypadków stanowi estymowaną moc testu.

```r
power <- function(p_true, n, N = 100) {

  MC_exact <- 0
  MC_asymptotic <- 0

  for (i in 1:N) {
    X <- rbinom(1, n, p_true)
    res = test(X, n)
    MC_exact <- MC_exact + (res$H[1] == "H1")
    MC_asymptotic <- MC_asymptotic + (res$H[2] == "H1")
  }

  MC_exact <- MC_exact / N
  MC_asymptotic <- MC_asymptotic / N

  data.frame(method = c('Exact', 'Asymptotic'),
             power = c(MC_exact, MC_asymptotic))
}
```

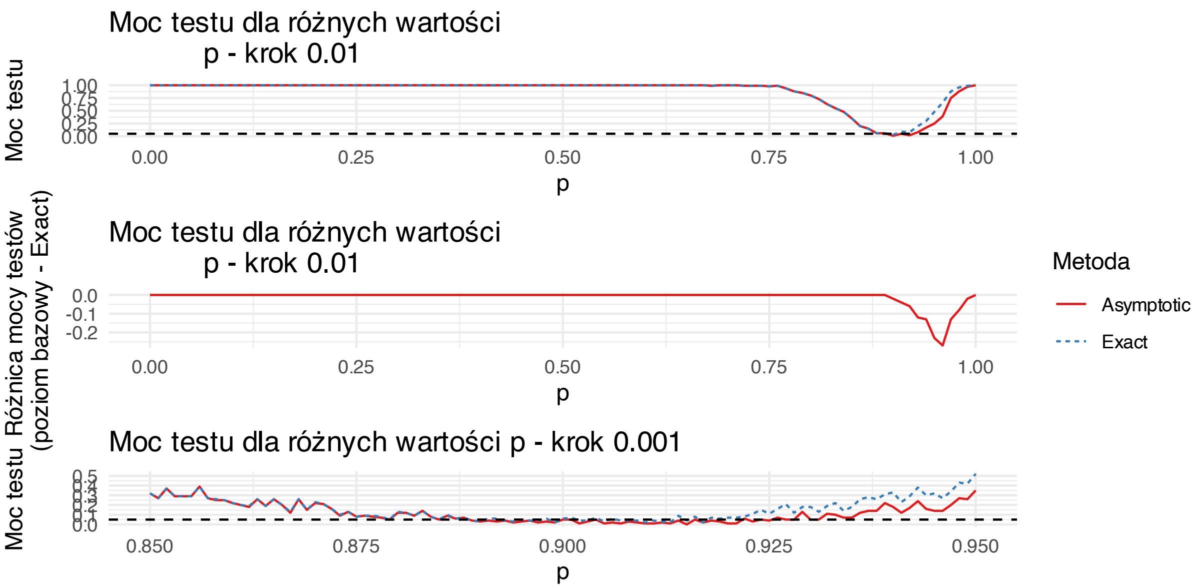
Kolejna funkcja, *plot_power*, pozwala wizualizować zarówno wartości mocy jako funkcję p, jak i różnicę mocy pomiędzy metodami, co pozwala na porównanie ich skuteczności.

```r
plot_power <- function(n, N = 100, ps = seq(0, 1, 0.01),
                       title="Moc testu dla różnych wartości p",
                       diff=FALSE) {

  plot_data <- data.frame()

  for (p in ps) {
    plot_data <- rbind(plot_data, cbind(data.frame(p=p), power(p, n, N)))
  }

  if (diff) {
    split_dfs <- split(plot_data, plot_data$method)

    exact_power <- split_dfs[["Exact"]][, "power"]

    split_dfs[["Asymptotic"]][, "power"] <-
      split_dfs[["Asymptotic"]][, "power"] - exact_power

    asymptotic_df <- split_dfs[["Asymptotic"]]
```

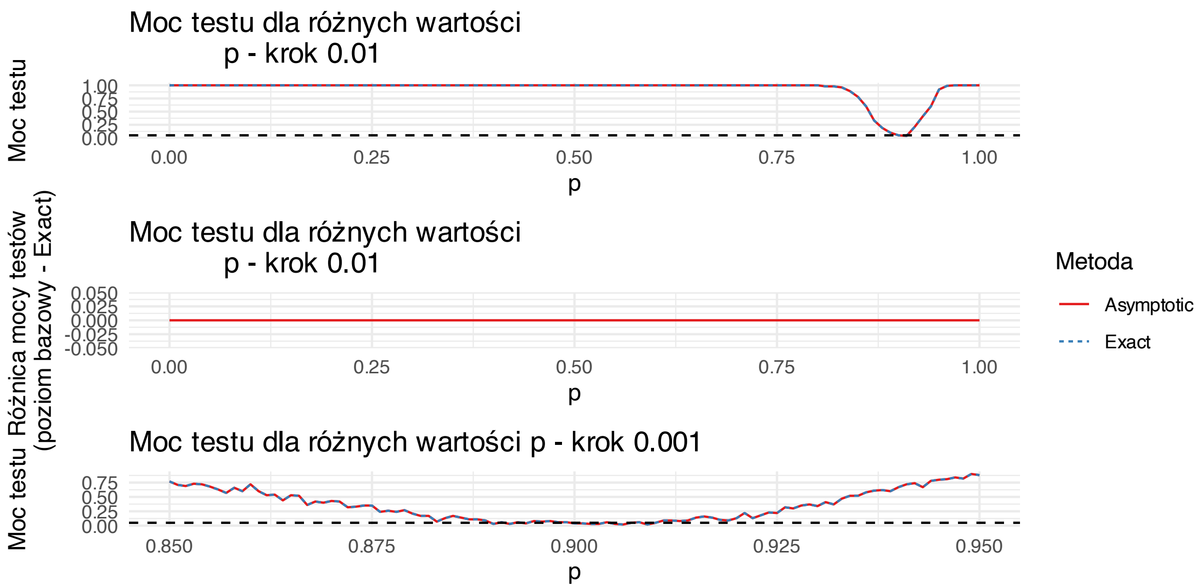
```
    p1 <- ggplot(
      plot_data, aes(x=p, y=power, color=method, linetype=method)
      ) +
      geom_line() +
      geom_hline(yintercept = 0.05, linetype="dashed") +
      scale_color_brewer(palette = "Set1", name = "Metoda") +
      scale_linetype_discrete(name = "Metoda") +
      labs(title = title,
           x = "p",
           y = "Moc testu") +
      theme_minimal() +
      theme(legend.position = "none")
    p2 <- ggplot(
      asymptotic_df, aes(x=p, y=power, color=method, linetype=method)
      ) +
      geom_line() +
      scale_color_brewer(palette = "Set1", name = "Metoda") +
      scale_linetype_discrete(name = "Metoda") +
      labs(title = title,
           x = "p",
           y = "Różnica mocy testów \n(poziom bazowy - Exact)") +
      theme(legend.position = "none") +
      theme_minimal()
    p1 / p2 + plot_layout(guides = 'collect') +
      theme(legend.position = "none")


  } else {
    ggplot(plot_data, aes(x=p, y=power, color=method, linetype=method)) +
      geom_line() +
      geom_hline(yintercept = 0.05, linetype="dashed") +
      scale_color_brewer(palette = "Set1", name = "Metoda") +
      scale_linetype_discrete(name = "Metoda") +
      labs(title = title,
           x = "p",
           y = "Moc testu") +
      theme_minimal()
  }
}
```

W dwóch ostatnich fragmentach kodu wykonano analizę dla rozmiarów próby odpowiednio $n = 100$ oraz $n = 300$, generując wykresy przedstawiające moc testów w zależności od wartości parametru p. Szczególne zainteresowanie poświęcono zakresom wartościom bliskim testowanej hipotezie zerowej.

```
n <- 100
plot_power(n, N = 100, title = 'Moc testu dla różnych wartości
           p - krok 0.01', diff=TRUE) /
  plot_power(n, N = 100, ps = seq(0.85, 0.95, 0.001),
             title = 'Moc testu dla różnych wartości p - krok 0.001') +
  plot_layout(guides = 'collect')
```




```
n <- 300
plot_power(n, N = 100, title = 'Moc testu dla różnych wartości
           p - krok 0.01', diff=TRUE) /
  plot_power(n, N = 100, ps = seq(0.85, 0.95, 0.001),
             title = 'Moc testu dla różnych wartości p - krok 0.001') +
  plot_layout(guides = 'collect')
```



Dla 100 kroków Monte-Carlo w przypadku rozmiaru próby $n = 100$ oraz $n = 300$ oba testy trzymały poziom istotności $\alpha = 0.05$ dla wartości p z hipotezy zerowej, tzm. moc jest poniżej poziomu istotności. W przypadku rozmiaru próby $n = 100$ moc testu dokładnego była wyższa od mocy testu asymptotycznego, co dobrze widać na wykresie różnicy między mocami testów, zatem test dokładny jest w tym wypadku testem jednostajnie najmocniejszym. W przypadku rozmiaru próby $n = 300$ moce obu testów są identyczne, dla tak dużych n kwestia wyboru testu nie ma większego znaczenia.

Zadanie dodatkowe

Wyznaczono granicę asymptotycznego przedziału ufności dla prawdopodobieństwa sukcesu bazując na przekształceniu logit korzystając z metody delta. Zaimplementowano metodę oraz porównano wyniki z funkcją zaimplementową w pakiecie *DescTools*.

W oparciu o funkcję centralną asymptotyczną

$$Q(Y,p) = \frac{\sqrt{n}(g(\hat{p}(Y) - g(p))}{g'(\hat{p}(Y))\hat{\sigma}}$$

dąży wg. rozkładu do $N(0,1)$ dla zmiennej losowej Y z rozkładu dwumianowego o liczbie prób n i nieznanym parametrze prawdopodbieństwa p. W celu wyznaczenia przedziału ufności logit wybieramy $g(p) = \log \frac{p}{1-p}$. Pochodna tej funkcji to $\frac{\mathrm{d}g}{\mathrm{d}p} = \frac{1}{p(1+p)}$. Za estymator odchylenia standardowego bierzemy $\hat{\sigma} = \sqrt{p(1-p)}$ i estymator p – estymator największej wiarygności $\hat{p} = \frac{Y}{n}$. Teraz wyznaczamy przedział ufności dla parmetru $g(p)$.

$$
\begin{aligned}
&P(-z_{1-\frac{\alpha}{2}} \leq Q(Y,p) \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha \\
&P(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(g(\widehat{p})-g(p))}{\frac{1}{\widehat{p}(1-\widehat{p})}\sqrt{\widehat{p}(1-\widehat{p})}} \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha \\
&P(g(\widehat{p}) - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\widehat{p}(1-\widehat{p})n}} \leq g(p) \leq g(\widehat{p}) + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\widehat{p}(1-\widehat{p})n}}) = 1-\alpha
\end{aligned}
$$

Zatem przdział ten jest postaci $[T_L', T_U']$, gdzie

$$
\begin{aligned}
T_L' &= g(\widehat{p}) - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\widehat{p}(1-\widehat{p})n}}, \\
T_U' &= g(\widehat{p}) + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\widehat{p}(1-\widehat{p})n}},
\end{aligned}
$$

gdzie z_β to kwantyl standardowego rozkładu normalnego rzędu β.

Biorąc funkcję odwrotną otrzymujemy asymptotyczny przedział ufności logit dla parametru p postaci $[T_L, T_U]$, gdzie

$$
\begin{aligned}
T_L &= g^{-1}\left(g(\widehat{p}) - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\widehat{p}(1-\widehat{p})n}}\right), \\
T_U &= g^{-1}\left(g(\widehat{p}) + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{\widehat{p}(1-\widehat{p})n}}\right).
\end{aligned}
$$

Poniżej zaimplementowano jego realizację dla obserwowanych realizacji zmiennych losowych.

```
logit_CI <- function(x, n, alpha=0.05) {
  p <- x / n
  se <- sqrt(1 / (p * (1 - p) * n))
  z <- qnorm(1 - alpha / 2)
  L <- Logit(p) - z * se
  U <- Logit(p) + z * se
  data.frame(est=p, lwr.ci=LogitInv(L), upr.ci=LogitInv(U))
}
```

Dla przykładowej realizacji zmiennej losowej z rozkładu $B(n,p)$, gdzie $n = 10$ oraz $p = 0.3$, obliczamy przedziały ufności obiema metodami.

```
x <- rbinom(1, 10, 0.3)

own <- logit_CI(x,10)
dtools <- BinomCI(x, 10, method="logit")
own
```

```
  est   lwr.ci    upr.ci
1 0.4 0.158342 0.7025951
```

```r
dtools
```

```
     est   lwr.ci    upr.ci
[1,] 0.4 0.158342 0.7025951
```

```r
atol <- 1e-8
all(abs(own[,2:3] - dtools[,2:3]) <= atol)
```

```
[1] TRUE
```

Wyniki są zgodne z tolerancją 10^{-8}, co potwierdza poprawność implementacji.