Analiza danych ankietowych

Sprawozdanie 1

Zuzanna Nasiłowska Maria Nowacka

Spis treści

1 Część 1

1.1 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).

1.1.1 Zadanie 1.1

Wczytamy dane i sprawdzimy ich rozmiar.

```
[1] 200   8
```

Dane zawierają 200 wierszy oraz 8 kolumn.

Sprawdzamy typy zmiennych.

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

Wszystkie zmienne o typie *character* przekształcamy na typ *factor*.

```
Liczba wartości brakujących wynosi: 0
```

Sprawdzamy, czy typy zmiennych zostały prawidłowo rozpoznane.

1. Zmienne ilościowe (typ numeric)

```
 STAŻ PYT_1 PYT_2 PYT_3  WIEK
    2     4     5     6     8
```

2. Zmienne jakościowe (typ factor)

```
  DZIAŁ CZY_KIER     PŁEĆ
     1        3        7
```

1.1.2 Zadanie 1.2

Utwórz zmienną WIEK_KAT przeprowadzając kategoryzacją zmiennej WIEK korzystając z nastąpujących przedziałów: do 35 lat, między 36 a 45 lat, między 46 a 55 lat, powyżej 55 lat.

1.1.3 Zadanie 1.3

Sporządź tablice liczności dla zmiennych: DZIAŁ, STAŻ, CZY_KIER, PŁEĆ, WIEK_KAT. Sformułuj wnioski.

```
DZIAŁ
HR IT MK PD
31 26 45 98

STAŻ
  1   2   3
 41 140  19

CZY_KIER
Nie Tak
173  27

PŁEĆ
  K   M
 71 129

WIEK_KAT
  <35 36-40 46-55   >55
   26   104    45    25
```

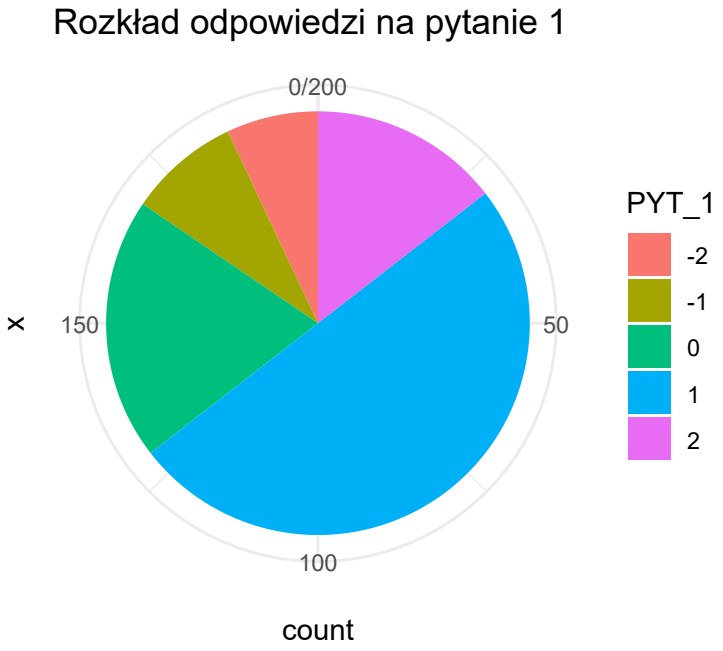
Na podstawie tabel liczności możemy zauważyć, że:

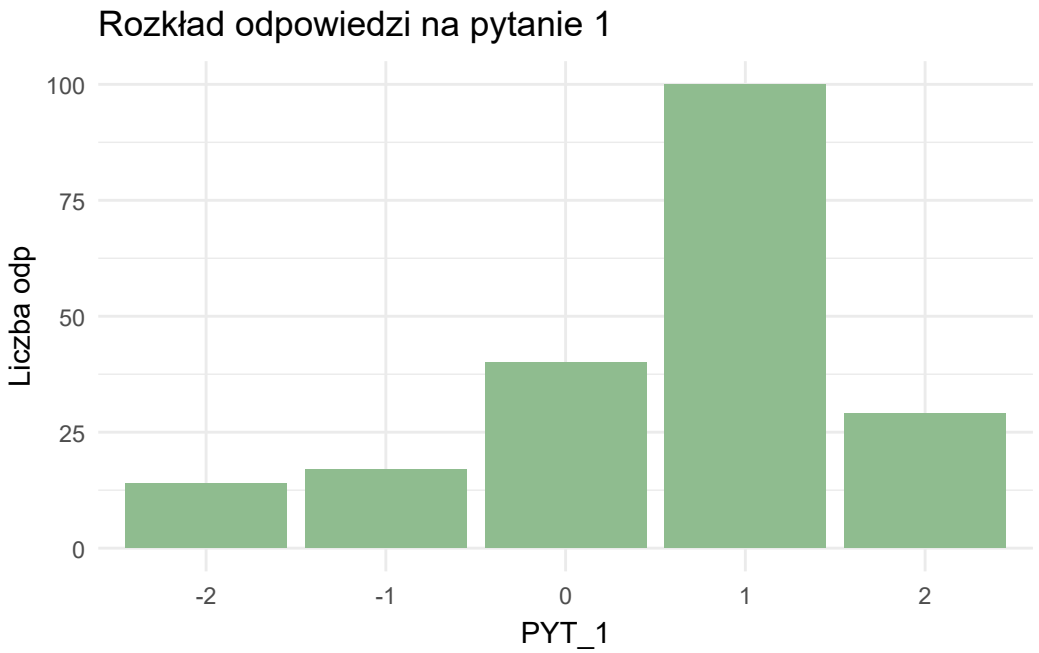
- W firmie prawie połowa pracowników jest zatrudniona w dziale **"PD"** (Dział Produktowy). Drugi największy dział to "MK" (Marketing), następnie "HR" (Dział zasobów ludzkich). Najmniej pracowników jest zatrudnionych w dziale "IT".
- Najwięcej osób pracuje w firmie między jednym a trzema latami. Mało osób ma staż ponad 3 lata.
- W firmie 27 osób ma stanowisko kierownicze (zdecydowana mniejszość)

- Większość pracowników to **mężczyźni**.
- Ponad połowa pracowników jest w wieku **36-40 lat**.

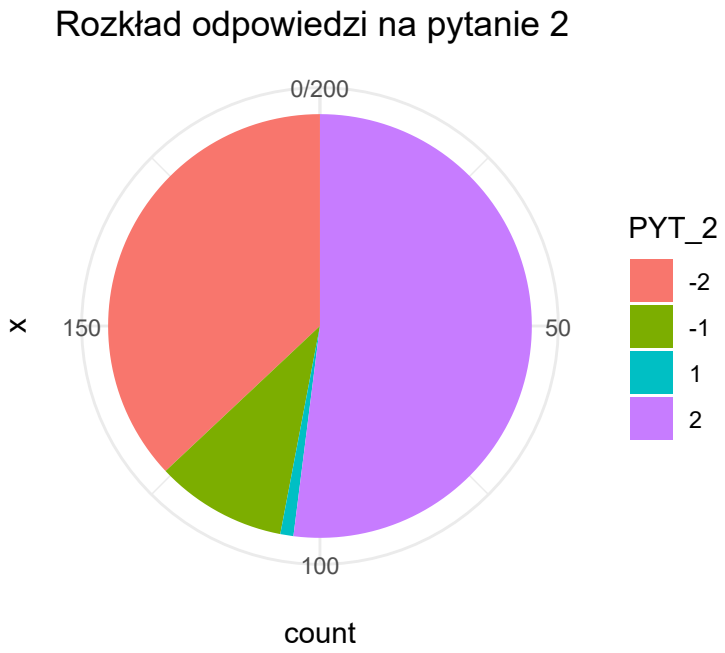
1.1.4 Zadanie 1.4

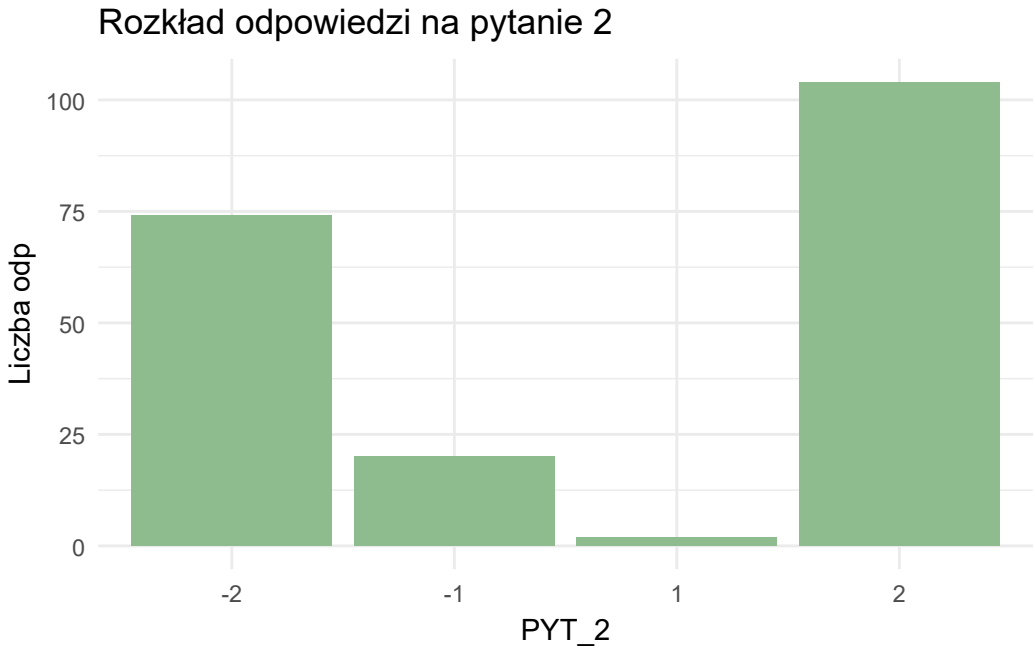
Sporządź wykresy kołowe oraz wykresy słupkowe dla zmiennych: PYT_1 oraz PYT_2. Sformułuj wnioski.





Pytanie 1 brzmiało: "Jak bardzo zgadzasz się ze stwierdzeniem, że firma zapewnia odpowiednie wsparcie i materiały umożliwiające skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń?" większość ankietowanych odpowiedziała 1 - "Zgadzam się" lub 2 - "Zdecydowanie się zgadzam". Prawie 1/4 osób nie ma zdania na ten temat. Możemy więc wnioskować, że więszość firmy jest zadowolona z przeprowadzanych szkoleń.





Na **pytanie 2**, o treści "Jak bardzo zgadzasz się ze stwierdzeniem, że firma oferuje szkolenia dostosowane do twoich potrzeb, wspierając twój rozwój zawodowy i szanse na awans?" nieco ponad połowa osób odpowiedziała "Zdecydowanie się zgadzam", jednak prawie wszyscy inni pracownicy dali odpowiedź "Nie zgadzam się" lub "Zdecydowanie się nie zgadzam", z przewagą

tych drugich. Na to pytanie pracownicy udzielili bardzo skrajnych odpowiedzi. Pomimo zadowolenia połowy pracowników, warto zbadać ten temat głębiej i przeprowadzić szkolenia dla tych, którzy nie czują się odpowiednio wspierani przez firmę.

1.1.5 Zadanie 1.5

Sporządź tablice wielodzielcze dla par zmiennych: PYT_1 i DZIAŁ, PYT_1 i STAŻ, PYT_1 i CZY_KIER, PYT_1i PŁEĆ C oraz PYT_1 i WIEK_KAT. Sformułuj wnioski.

```
     DZIAŁ
PYT_1 HR IT MK PD
   -2  2  0  3  9
   -1  2  2  3 10
   0   5  4 14 17
   1  19 15 15 51
   2   3  5 10 11
```

```
     STAŻ
PYT_1  1  2  3
   -2  5  5  4
   -1  6 10  1
   0   8 26  6
   1  19 75  6
   2   3 24  2
```

```
     CZY_KIER
PYT_1 Nie Tak
   -2  10   4
   -1  14   3
   0   34   6
   1   88  12
   2   27   2
```

```
     PŁEĆ
PYT_1  K  M
   -2  3 11
   -1  7 10
   0  14 26
   1  36 64
   2  11 18
```

```
     WIEK_KAT
PYT_1 <35 36-40 46-55 >55
   -2   1    11     2   0
   -1   6     7     1   3
   0    3    24     5   8
   1   13    50    25  12
   2    3    12    12   2
```

Wnioski (*to jeszcze jakoś ładniej ująć w słowa*)

zadowolenie = zgadza się z stwierdzeniem

- dział:
  - najwięcej niezadowolonych osób jest w dziale PD ale to największy dział
  - IT wydaje się być w większości zadowolony
- staż:
  - dla osób z niższym stażem około połowa osób jest zadowolona, reszta nie ma zdania lub jest niezadowolona.
  - dla osob ze stażem miedzy 1 a 3 lata mamy bardzo dużą grupę osób zadowolonych, jednak całkiem sporo osób zaznaczyło opcję “nie mam zdania”.
- kierownictwo
  - około 1/4 kierowników jest niezadowolona.
  - Dla nie-kierowników odpowiedzi rozkładają się bardziej w kierunku pozytywnym
- płeć:
  - kobiety są bardziej zadowolone (procentowo)
- wiek:
  - największy odsetek niezadowolonych osób jest wśród najmłodszych pracowników a najmniejszy w grupie 46-55 lat

1.1.6 Zadanie 1.6

Sporządź tablicę wielodzielczą dla pary zmiennych: PYT_2 i PYT_3. Sformułuj wnioski.

```
     PYT_3
PYT_2 -2 -1  1  2
   -2 49 16  5  4
   -1  3  6 10  1
   1   0  0  2  0
   2   0  8 15 81
```

Wnioski

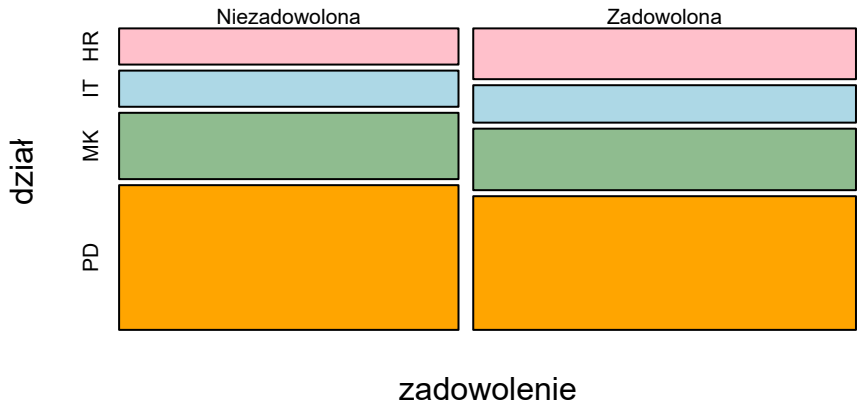
Duże grupy osób zostały przy swojej silnej opini (-2 i 2). Sumarycznie około 15% głosów zmieniono na bardziej pozytywne, jednak w ponad 10% przypadków opinia zmieniła się na gorszą. Sugeruje to, że część osób odczuła pozytywne skutki szkoleń, jednak nadal pozostaje grupa osób, którym one nie pomogły, a nawet zaszkodziły.

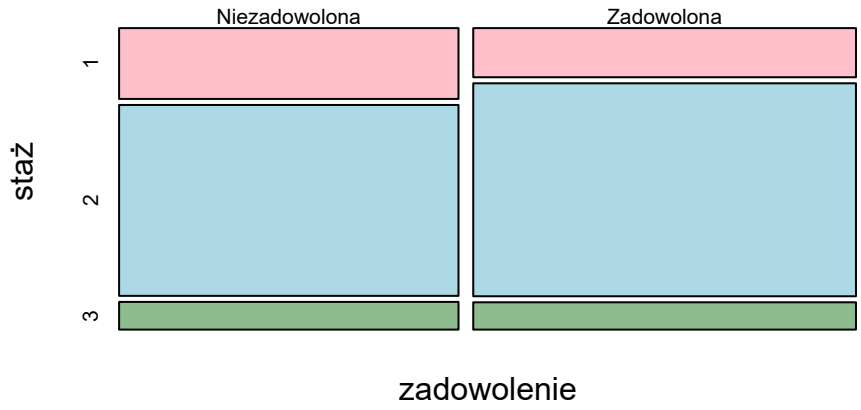
1.1.7 Zadanie 1.7

Utwórz 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".

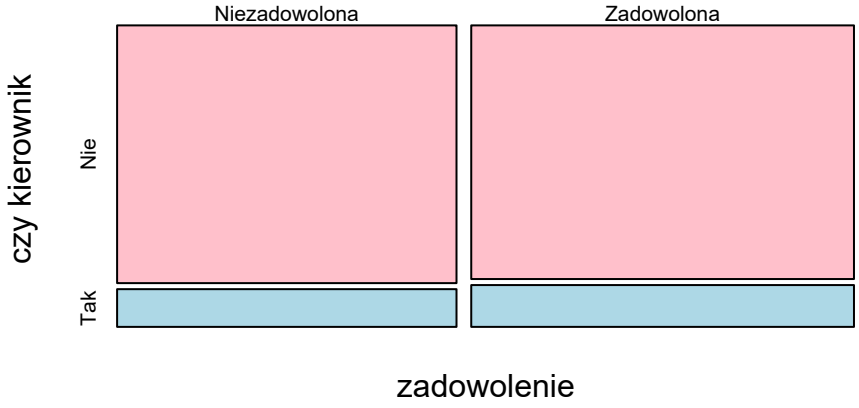
1.1.8 Zadanie 1.8

Sporządź 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. Czy na podstawie uzyskanch wykresów można postawić pewne hipotezy dotyczące realicji między powyższymi zmiennymi? Spróbuj sformułować kilka takich hipotez.

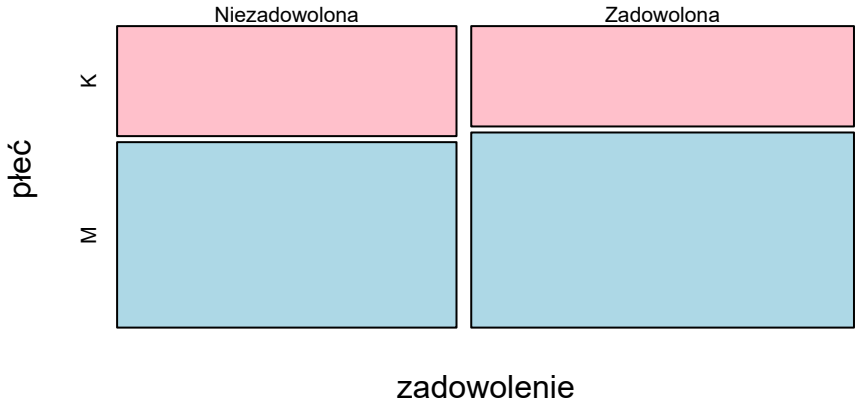

zadowolenie z podziałem na działy
Niezadowolona
Zadowolona
dział
HR
IT
MK
PD
zadowolenie

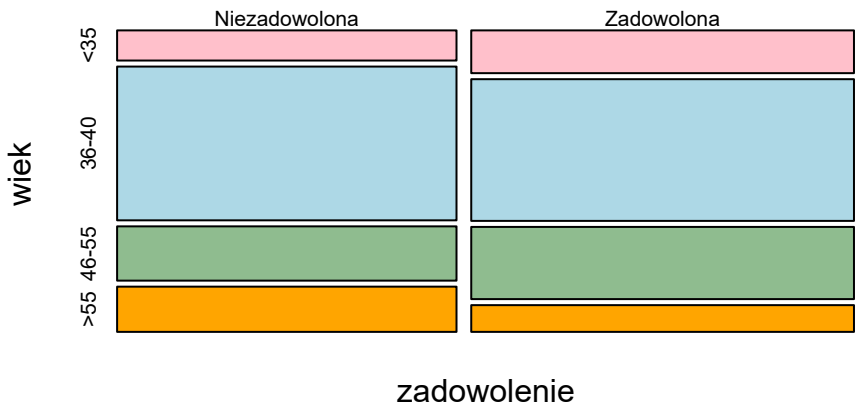

zadowolenie z podziałem na staż
Niezadowolona
Zadowolona
staż
1
2
3
zadowolenie

zadowolenie z podziałem na kierownictwo i resztę



zadowolenie z podziałem na płeć





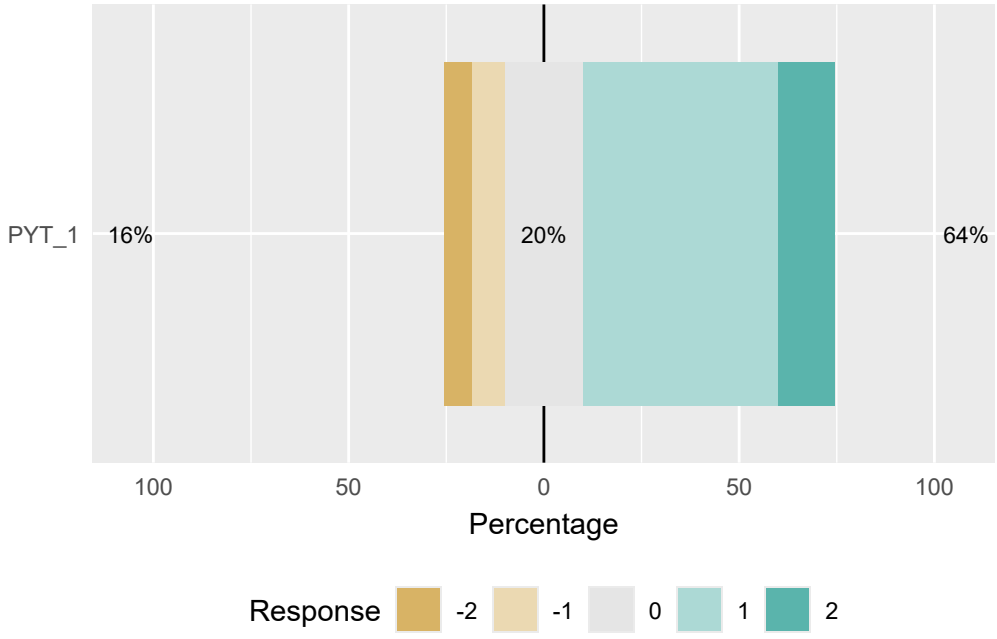
Badając odpowiedzi na **pytanie 2**, przy podziale pracowników na odpowiednie grupy możemy zauważyć:

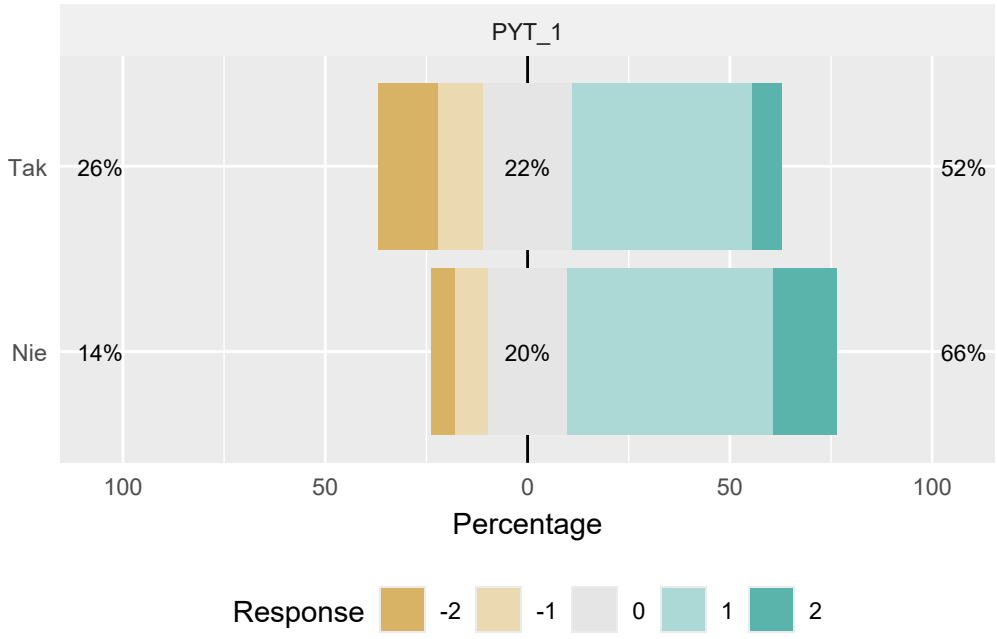
- DZIAŁ: widzimy, że dla działu "PD" oraz "MK" więcej jest osób niezadowolonych, a w dziale "HR" więcej mamy osób zadowolonych. W dziale "IT" jest mniej więcej tyle samo zadowolonych i niezadowolonych osób. Widzimy zależność między badanymi zmiennymi.
- STAŻ: osoby o najmniejszym stażu są w większości niezadowolone, Dla grupy 1-3 widzimy zadowolenie większości, a w ostatniej grupie odpowiedzi rozkładają się po równo. Moglibyśmy przetestować jeszcze raz tę zależność dla bardziej szczegółowego podziału osób według długości stażu, teraz widzimy niezbyt silną korelację.
- CZY_KIER: przy tym podziale nie widać drastycznych nierówności. Osoby o stanowisku kierowniczym są delikatnie częściej zadowolone od pozostałych. Nie widać jednak silnej zależności między tymi zmiennymi.
- PŁEĆ: wiecej kobiet jest niezadowolonych, a w grupie mężczyzn delikatnie przeważają osoby zadowolone. Ponownie nie widać silnej zależności.
- WIEK_KAT: w grupach "36-40" oraz ">55" przeważają odpowiedzi negatywne (niezadowolenie), a w pozostałych - pozytywne. Widzimy tutaj pewną nieliniową zależność.

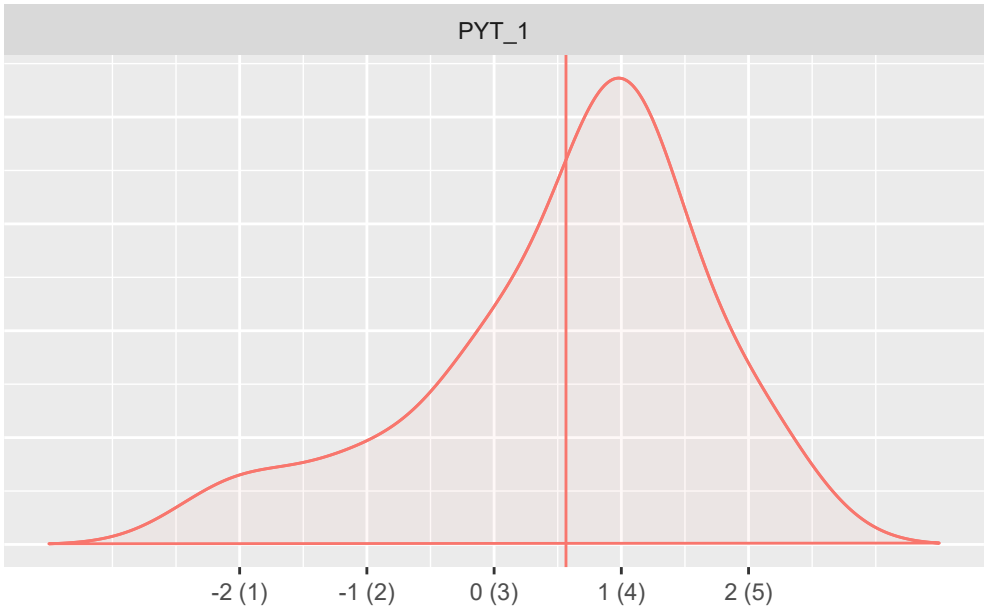
2 Część 2

2.1 Zadanie 2

Zilustruj 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. W tym celu możesz zaproponować własne metody wizualizacji lub zapozać sić z bibliotekć `likert` i dostępnymi tam funkcjami `summary` oraz `plot` (jeśli korzystarz z R) oraz z bibliotek `Altair` lub `plot-likert` (jeśli korzystarz z Pythona).







Na pierwszym i ostatnim wykresie widzimy przewagę odpowiedzi “1” i “2”, nad pozostałymi “-2”, “-1” i “0”. Jednak po podzieleniu grupy badanych ze względu na zmienną CZY_KIER widzimy większe niezodowolenie w grupie kierowników. Osoby bez stanowisk kierowniczych rzadziej udzielały negatywnych odpowiedzi i częściej głosowały na opcję “Zdecydowanie się

zgadzam".

2.2 Zadanie 3

Zapoznaj się z funkcją `sample` z biblioteki `stats` (w R) lub z funkcją `random.choice` z biblioteki numpy (w Pythonie). Przetestuj jej działanie dla różnych wartości argumentów wejściowych. Następnie wylosuj próbkę o liczności 10% wszystkich rekordów z pliku "ankieta.csv" w dwóch wersjach: ze zwracaniem oraz bez zwracania

```
library(stats)
bez_zwracania <- ankieta[sample(1:nrow(ankieta), size = 0.1*nrow(ankieta), replace = FALSE),]
ze_zwracaniem <- ankieta[sample(1:nrow(ankieta), size = 0.1*nrow(ankieta), replace = TRUE),]
```

2.3 Zadanie 4

Zaproponuj metodę symulowania zmiennych losowych z rozkładu dwumianowego. Napisz funkcję do generowania realizacji, a następnie zaprezentuj jej działanie porównujęc wybrane teoretyczne i empiryczne charakterystyki dla przykładowych wartości paramertów rozkładu: n i p.

```
symulacja <- function(N,n, p) {

  wyniki <- numeric(N)

  for(i in 1:N) {
    bernoulli <- rbinom(n = n, size = 1, prob = p)
    wyniki[i] <- sum(bernoulli)
  }

  return(wyniki)
}
n <- 200
p <- 0.2
N <- 10000
```

```
Teoretyczna wartość oczekiwana:  40

Teoretyczna wariancja:  32

empiryczna wartość oczekiwana:  40.0305

empiryczna wariancja:  32.95847
```

2.4 Zadanie 5

Zaproponuj metodę symulowania wektorów losowych z rozkładu wielomianowego. Napisz funkcję do generowania realizacji, a następnie zaprezentuj jej działanie porównujęc wybrane teoretyczne i empiryczne charakterystyki dla przykładowych wartości paramertów rozkładu: n i **p**.

```
los_wiel <- function(ps, N){
  k <- length(ps)
  csum = cumsum(ps)
  X <- rep(0, k)
  for (i in 1:N){
    Z <- runif(1)
    for (j in 1:k){
      if (Z<csum[j]){
      X[j] <- X[j] + 1
      break }
    }
  }
  return(X/N)
}
```

```
Podany wektor prowdopodobieństwa:     0.1 0.23 0.47 0.17 0.03
```

```
Empiryczny rozkład prawdopodobieństwa 0.1037 0.2336 0.4633 0.1683 0.0311
```

3 Część 3

3.1 Zadanie 6

Napisz funkcję do wyznaczania realizacji przedziału ufności Cloppera-Pearsona. Niech argumentem wejściowym będzie poziom ufności, liczba sukcesów i liczba prób lub poziom ufności i wektor danych (funkcja powinna obsługiwać oba przypadki).

```
clopper_pearson <- function(alpha, sukces, n = NULL){
  if(is.null(n)){
    data <-  sukces
    sukces <- sum(data == "1")
    n <- length(data)
```

```
  }
  if(sukces == 0){
    p_dol <-  0
  } else{
    p_dol <-  qbeta(alpha, sukces, n-sukces - 1)
  }
  if(sukces == n){
    p_gora <-  1
  } else{
    p_gora <- qbeta(alpha, sukces + 1, n - sukces)
  }
  return(c(p_dol, p_gora))
}
```

3.2 Zadanie 7

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

```
Przedział dla zmiennej 'CZY_ZADOW':  0.4583305 0.6007671
```

```
Przedział dla zmiennej 'CZY_ZADOW2': 0.5184216 0.6588694
```

3.3 Zadanie 8

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

```
[1] 368 367 370 367 369
```

```
          method x  n      mean       lower     upper
1  agresti-coull 2 10 0.2000000  0.04588727 0.5206324
2  agresti-coull 4 10 0.4000000  0.16711063 0.6883959
3     asymptotic 2 10 0.2000000 -0.04791801 0.4479180
4     asymptotic 4 10 0.4000000  0.09636369 0.7036363
5          bayes 2 10 0.2272727  0.02346550 0.4618984
```

```
6          bayes 4 10 0.4090909  0.14256735 0.6838697
7        cloglog 2 10 0.2000000  0.03090902 0.4747147
8        cloglog 4 10 0.4000000  0.12269317 0.6702046
9          exact 2 10 0.2000000  0.02521073 0.5560955
10         exact 4 10 0.4000000  0.12155226 0.7376219
11         logit 2 10 0.2000000  0.05041281 0.5407080
12         logit 4 10 0.4000000  0.15834201 0.7025951
13        probit 2 10 0.2000000  0.04206918 0.5175162
14        probit 4 10 0.4000000  0.14933907 0.7028372
15       profile 2 10 0.2000000  0.03711199 0.4994288
16       profile 4 10 0.4000000  0.14570633 0.6999845
17           lrt 2 10 0.2000000  0.03636544 0.4994445
18           lrt 4 10 0.4000000  0.14564246 0.7000216
19     prop.test 2 10 0.2000000  0.03542694 0.5578186
20     prop.test 4 10 0.4000000  0.13693056 0.7263303
21        wilson 2 10 0.2000000  0.05668215 0.5098375
22        wilson 4 10 0.4000000  0.16818033 0.6873262
```

3.4 Zadanie 9

```r
library(binom)
n_values <- c(30, 100, 1000)
p_values <- seq(0.01, 0.99, by=0.01)
N <- 50
metody <- c('exact', 'asymptotic', 'wilson')
```

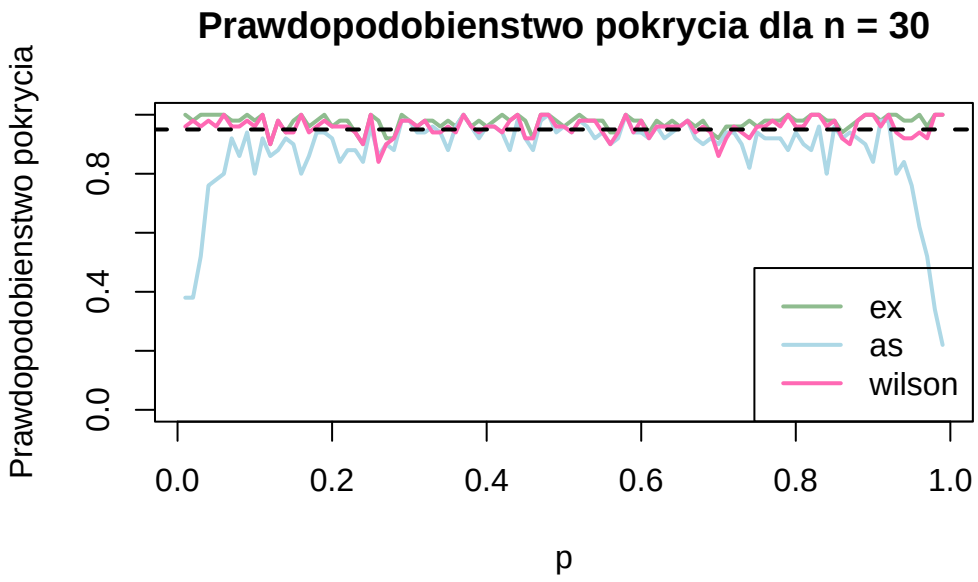
```r
simulate_confidence_intervals <- function(metody, n, p_values, alpha = 0.05, N = 500){
  coverage_results <- matrix(NA, nrow = length(metody), ncol = length(p_values))
  length_results <- matrix(NA, nrow = length(metody), ncol = length(p_values))

  for (j in 1:length(p_values)) {
    p <- p_values[j]

    coverage_ex <- numeric(N)
    coverage_as <- numeric(N)
    coverage_wilson <- numeric(N)

    length_ex <- numeric(N)
    length_as <- numeric(N)
    length_wilson <- numeric(N)
```

```
    for (t in 1:N) {
      x <- rbinom(1, n, p)

      exact <- binom.confint(x, n, conf.level = 1 - alpha, method = "exact")
      as <- binom.confint(x, n, conf.level = 1 - alpha, method = "asymptotic")
      wilson <- binom.confint(x, n, conf.level = 1 - alpha, method = "wilson")

      coverage_ex[t] <- (p >= exact$lower && p <= exact$upper)
      coverage_as[t] <- (p >= as$lower && p <= as$upper)
      coverage_wilson[t] <- (p >= wilson$lower && p <= wilson$upper)

      length_ex[t] <- exact$upper - exact$lower
      length_as[t] <- as$upper - as$lower
      length_wilson[t] <- wilson$upper - wilson$lower
    }

    coverage_results[1, j] <- mean(coverage_ex)
    length_results[1, j] <- mean(length_ex)
    coverage_results[2, j] <- mean(coverage_as)
    length_results[2, j] <- mean(length_as)
    coverage_results[3, j] <- mean(coverage_wilson)
    length_results[3, j] <- mean(length_wilson)

    }

  list(coverage = coverage_results, length = length_results)
}

results30 <- simulate_confidence_intervals(metody, 30, p_values, N = N)
results100 <- simulate_confidence_intervals(metody, 100, p_values, N = N)
results1000 <- simulate_confidence_intervals(metody, 1000, p_values, N = N)
```

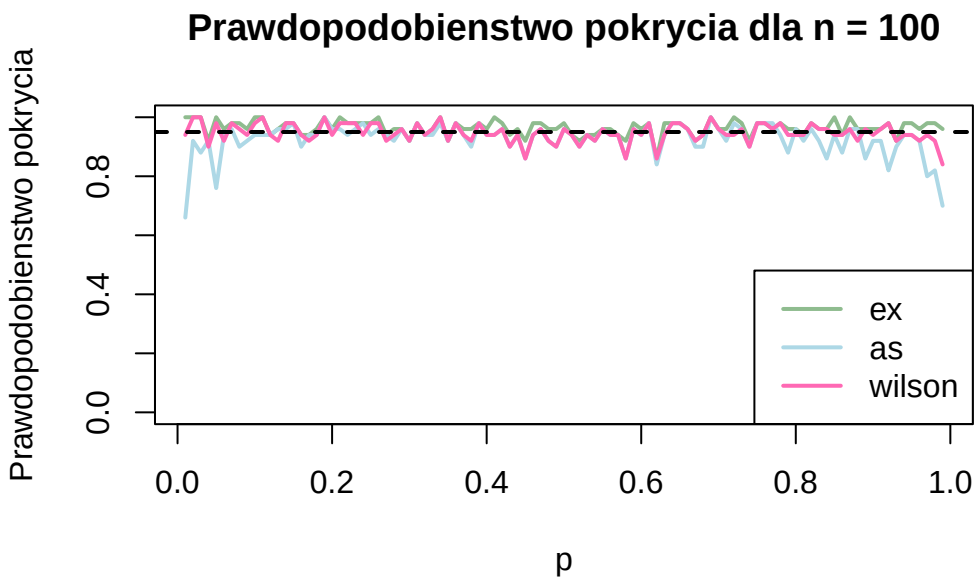
```
my_colors <- c('darkseagreen', 'lightblue', 'hotpink')
plotowanie <- function(p_values, results, tit1, tit2){
    plot(p_values, results[1,], type = "l", col = my_colors[1], lwd = 2,
     xlab = 'p', ylab = tit2, , ylim = c(0, max(results)),
     main = tit1)
lines(p_values, results[2,], col = my_colors[2], lwd = 2)
lines(p_values, results[3,], col = my_colors[3], lwd = 2)
legend("bottomright", legend = c("ex", "as", "wilson"),
       col = my_colors, lwd = 2)
if (grepl("Prawdopodobieństwo pokrycia", tit1)) {
```

```
abline(h = 0.95, col = "black", lwd = 2, lty = 2)
}}
```

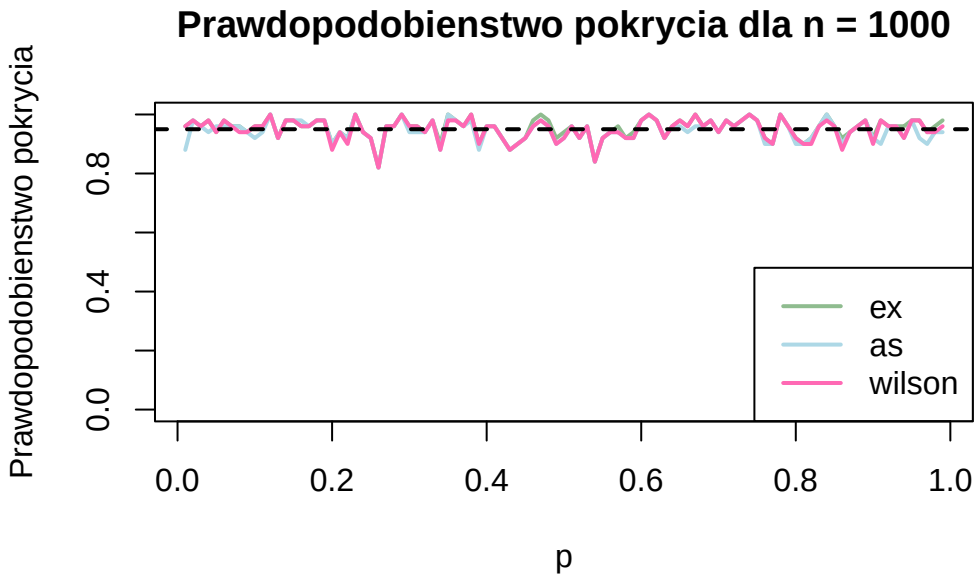
```
plotowanie(p_values, results30$coverage, "Prawdopodobieństwo pokrycia dla n = 30", "Prawdopo
```



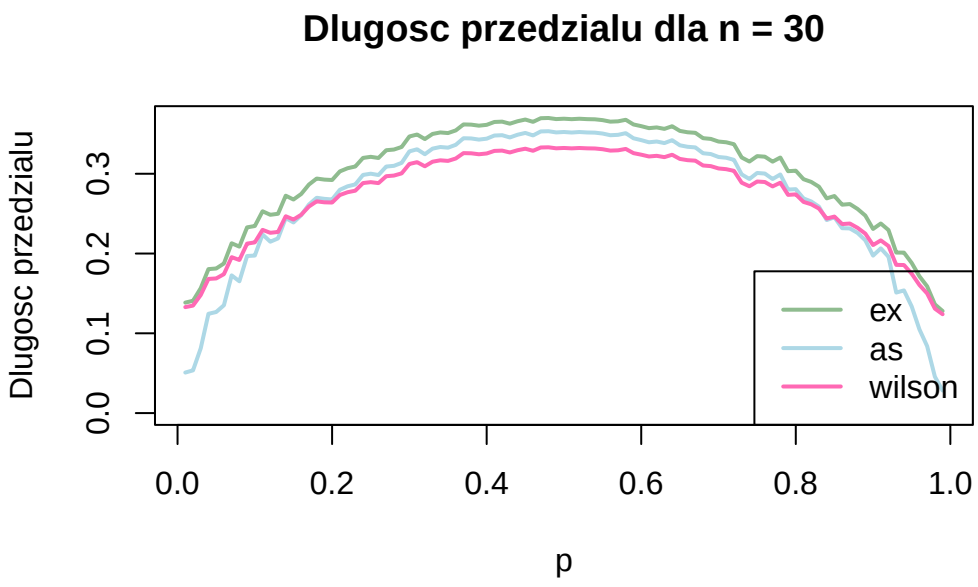

```
plotowanie(p_values, results100$coverage, "Prawdopodobieństwo pokrycia dla n = 100", "Prawdo
```



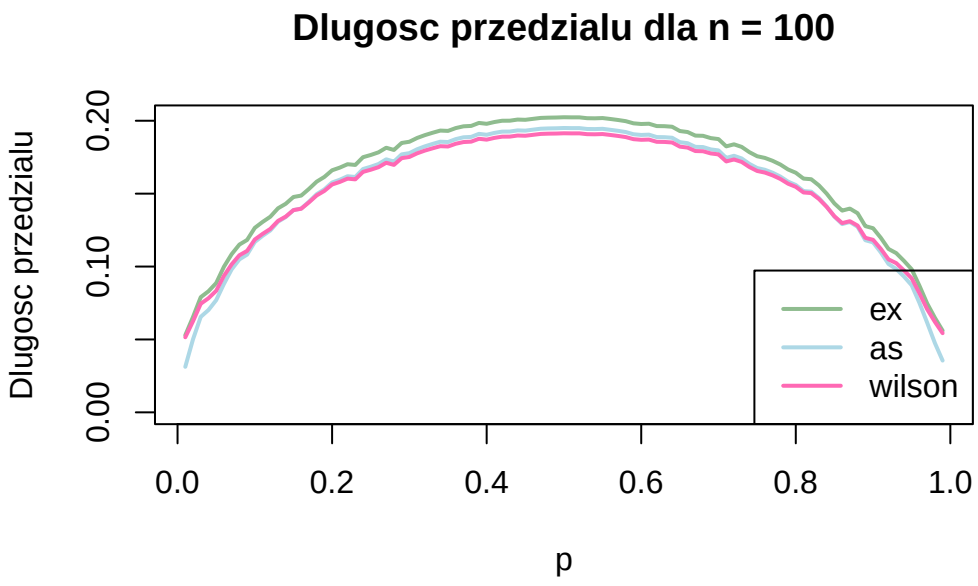

```
plotowanie(p_values, results1000$coverage, "Prawdopodobieństwo pokrycia dla n = 1000", "Prawd
```



```
tit2 = 'Długość przedziału'
plotowanie(p_values, results30$length, "Długość przedziału dla n = 30", tit2)
```




```
plotowanie(p_values, results100$length, "Długość przedziału dla n = 100", tit2)
```




```
plotowanie(p_values, results1000$length, "Długość przedziału dla n = 1000", tit2)
```