Analiza danych ankietowych - sprawozdanie nr 1

Bartłomiej Gintowt

30.03.2023

1 Wstęp

Celem naszej pracy będzie przygotowanie sprawozdania, w którym dokonamy analizy otrzymanych danych ankietowych uzyskanych podczas badania losowo wybranych dwustu pracowników wielkiej korporacji. Wykorzystane zostało losowanie proste ze zwracaniem. Dodatkowo zaaplikujemy potrzebne do tego algorytmy oraz napisane przez nas funkcje.

2 Wczytanie i objaśnienie danych

Pierwszym zadaniem, którego się podejmiemy, będzie odpowiednie wczytanie danych z pliku csv oraz nadanie kolumnom odpowiednich nazw. Poniżej znajduje się kod, za pomocą którego wczytujemy dane, a na rysunku umieściliśmy kilka przykładowych rekordów.

```
dane <- read.csv2("personel.csv",
col.names = c("D", "S", "A1", "A2", "W1", "W2", "P", "Wiek", "Wyk"))
```

	D	S	A1	A2	W1	W2	P	Wiek	Wyk
1	O	0	1	1	-2	-2	M	4	2
2	O	0	0	0	-2	-2	M	4	2
3	O	0	1	1	2	2	M	4	2
4	O	0	-1	0	-2	-2	K	4	2
5	O	1	1	1	2	2	K	4	3
6	O	1	0	0	1	2	K	4	3
7	O	1	2	2	2	2	K	4	2

Rysunek 1: Przykładowe pierwsze siedem wierszy wczytanych danych.

Kolumna D odpowiada za dział, w którym pracuje dany pracownik. Możemy znaleźc w niej 4 różne wartości: Z (dział zaopatrzenia), P (dział produkcyjny), S (dział sprzedaży) i O (dział obsługi kadrowo-płacowej).

Zmienna S określa, czy dany pracownik piastuje stanowisko kierownicze. Przyjmuje ona dwie wartości: 0, jeżeli tak nie jest, oraz 1 w przeciwnej sytuacji.

W kolumnach A1 i A2 znajdują się oceny pracowników nt. stwierdzenia, że atmosfera panująca w pracy jest bardzo dobra. Dane te zebrano za pomocą skali Likerta: możliwe odpowiedzi były liczbami całkowitymi od -2 (zdecydowanie się nie zgadzam) do 2 (zdecydowanie się zgadzam). Wyniki tego pytania zostały zanotowane w dwóch kolumnach, ponieważ zadano je dwukrotnie z roczną przerwą między nimi.

Kolejnym pytaniem w ankiecie była ocena nt. zadowolenia ze swojego wynagrodzenia. Tutaj, podobnie jak poprzednio, dwukrotnie zebrano opinie, a wyniki zostały umieszczone w kolumnach W1 i W2. Ankietowani udzielali odpowiedzi w skali podobnej do tej z pytania o atmosferze w pracy, ale bez opcji "trudno powiedzieć". Świadczy to o aplikacji skali Stapela do tego pytania.

Ostatnią częścią tej ankiety jest metryczka, w której znajdowały się pytania odnośnie płci (P), wieku (Wiek) i wykształcenia (Wyk) respondenta. Przyjmują one wartości:

- Płeć: K (kobieta), M (mężczyzna);
- Wiek: 1 (do 25 lat), 2 (26-35 lat), 3 (36-50 lat), 4 (powyżej 50 lat);

- Wykształcenie: 1 (zawodowe), 2 (średnie), 3 (wyższe).

Cała tabela składa się z 200 rekordów.

3 Część pierwsza

3.1 Zadanie 1

Wpierw sporządzimy tablice liczności dla zmiennych A1 oraz W1, biorąc pod uwagę wszystkie dane, jak i również te w podgrupach ze względu na zmienną dział, płeć oraz wykształcenie.

Tablicę liczności zmiennej A1 dla wszystkich danych utworzymy odpowiednio modelując nasze dane oraz wykorzystując bibliotekę xtable do utworzenia tablicy liczności.

```
library(xtable)
dane_a1 <- dane %>% group_by(A1) %>% count()
print(xtable(dane_a1 %>% t(),caption="Tabela liczności zmiennej A1"),
include.colnames = FALSE, table.placement = "H")
```

Otrzymujemy wówczas następującą tablice liczności.

A1	-2	-1	0	1	2
n	14	17	40	99	29

Tabela 1: Tablica liczności zmiennej A1.

Utworzymy teraz tablice liczności dla zmiennej A1 ze względu na płeć poprzez dodanie odpowiedniego czynnika modulującego przy pomocy wyrażenia filter(). Konstruujemy tablicę dla kobiet.

```
dane_k <- dane %>% filter(P=="K") %>% group_by(A1) %>% count()
```

A1	-2	-1	0	1	2
n	3	7	14	36	11

Tabela 2: Tablica liczności zmiennej A1 dla kobiet.

W bardzo podobny sposób utworzymy kolejne tablice.

A1	-2	-1	0	1	2
n	11	10	26	63	18

Tabela 3: Tablica liczności zmiennej A1 dla mężczyzn.

A1	-2	-1	0	1	2
n	2	2	5	19	3

Tabela 4: Tablica liczności zmiennej A1 dla działu zaopatrzenia

A1	-2	-1	0	1	2
n	9	10	17	51	11

Tabela 5: Tablica liczności zmiennej A1 dla działu produkcji

A1	-2	-1	0	1	2
n	3	3	14	15	10

Tabela 6: Tablica liczności zmiennej A1 dla działu sprzedaży

A1	-1	0	1	2
n	2	4	14	5

Tabela 7: Tablica liczności zmiennej A1 dla działu obsługi

A1	-2	-1	0	1	2
n	5	6	8	19	3

Tabela 8: Tablica liczności zmiennej A1 dla wykształcenia zawowodwego

A1	-2	-1	0	1	2
n	5	10	26	74	24

Tabela 9: Tablica liczności zmiennej A1 dla wykształcenia średniego

A1	-2	-1	0	1	2
n	4	1	6	6	2

Tabela 10: Tablica liczności zmiennej A1 dla wykształcenia wyższego

Pozostało nam utworzenie tablic liczności zmiennej W1, która bez podziału na podgrupy wygląda następująco.

W1	-2	-1	1	2
n	73	20	2	104

Tabela 11: Tabela liczności zmiennej W1

Uwzględniając odpowiednie podgrupy otrzymujemy następujące tablice liczności.

W1	-2	-1	1	2
n	25	10	1	35

Tabela 12: Tabela liczności zmiennej W1 dla kobiet

W1	-2	-1	1	2
n	48	10	1	69

Tabela 13: Tabela liczności zmiennej W1 dla mężczyzn

W1	-2	-1	2
n	9	3	19

Tabela 14: Tablica liczności zmiennej W1 dla działu zaopatrzenia

W1	-2	-1	1	2
n	37	11	1	49

Tabela 15: Tablica liczności zmiennej W1 dla działu produkcji

W1	-2	-1	2
n	20	2	23

Tabela 16: Tablica liczności zmiennej W1 dla działu sprzedaży

W1	-2	-1	1	2
n	7	4	1	13

Tabela 17: Tablica liczności zmiennej W1 dla działu obsługi

W1	-2	-1	2
n	20	3	18

Tabela 18: Tablica liczności zmiennej W1 dla wykształcenia zawowodwego

W1	-2	-1	2
n	44	17	78

Tabela 19: Tablica liczności zmiennej W1 dla wykształcenia średniego

W1	-2	1	2
n	9	2	8

Tabela 20: Tablica liczności zmiennej W1 dla wykształcenia wyższego

3.2 Zadanie 2

Teraz sporządzimy tabelę wielodzielczą uwzględniającą zmienne W1 i P, W1 i S oraz A1 i D. Wykorzystamy do tego funkcję structable. Zaczniemy od tabeli uwzględniającej zmienne W1 i P.

```
structable(W1~P,dane)
struc1 <- structable(W1~P, dane)
print(xtableFtable(ftable(struc1), method = "compact",
caption="Tabela wielodzielcza uwzględniająca zmienną W1 i P."))
```

P | W1	-2	-1	1	2
K	25	10	1	35
M	48	10	1	69

Tabela 21: Tabela wielodzielcza uwzględniająca zmienną W1 i P.

Podobnie utworzymy tabele wielodzielcze dla pozostałych par zmiennych.

S | W1	-2	-1	1	2
0	63	18	0	91
1	10	2	2	13

Tabela 22: Tabela wielodzielcza uwzględniająca zmienną W1 i S.

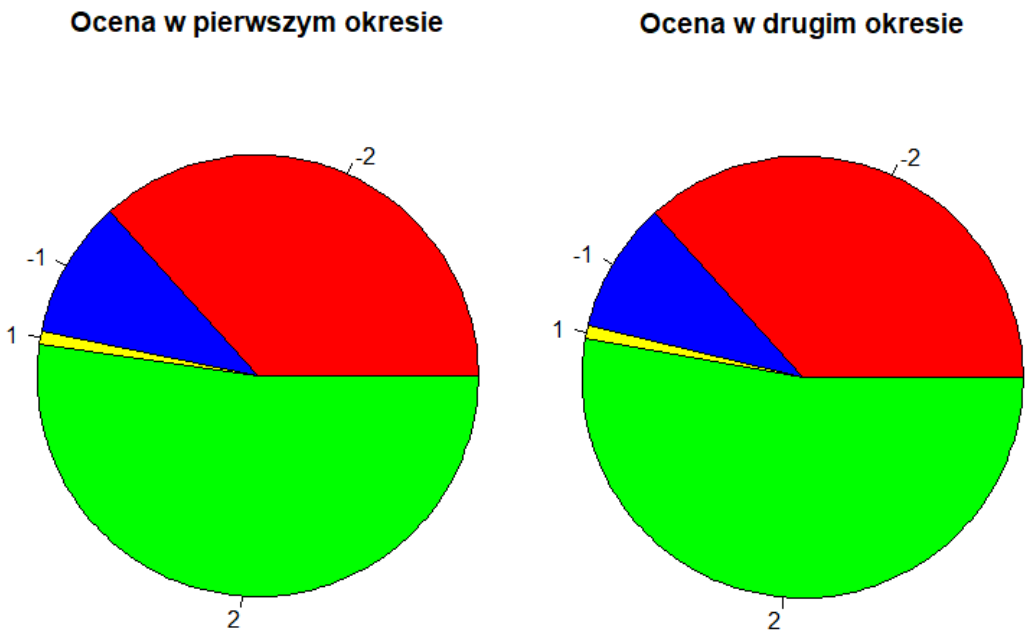
D | A1	-2	-1	0	1	2
O	0	2	4	14	5
P	9	10	17	51	11
S	3	3	14	15	10
Z	2	2	5	19	3

Tabela 23: Tabela wielodzielcza uwzględniająca zmienną A1 i D.

3.3 Zadanie 3

Naszym zadaniem będzie utworzenie wykresu kołowego oraz słupkowego dla zmiennej W1 i W2, gdzie W1 oznacza ocenę zadowolenia z zarobków pracowników w pierwszym okresie badania, a W2 w drugim okresie. Utworzymy wykresy kołowe wpierw dla zmiennej W1, a następnie dla W2.

```
pie(table(dane$W1), col=c("red", "blue", "yellow", "green"))
pie(table(dane$W2), col=c("red", "blue", "yellow", "green"))
```

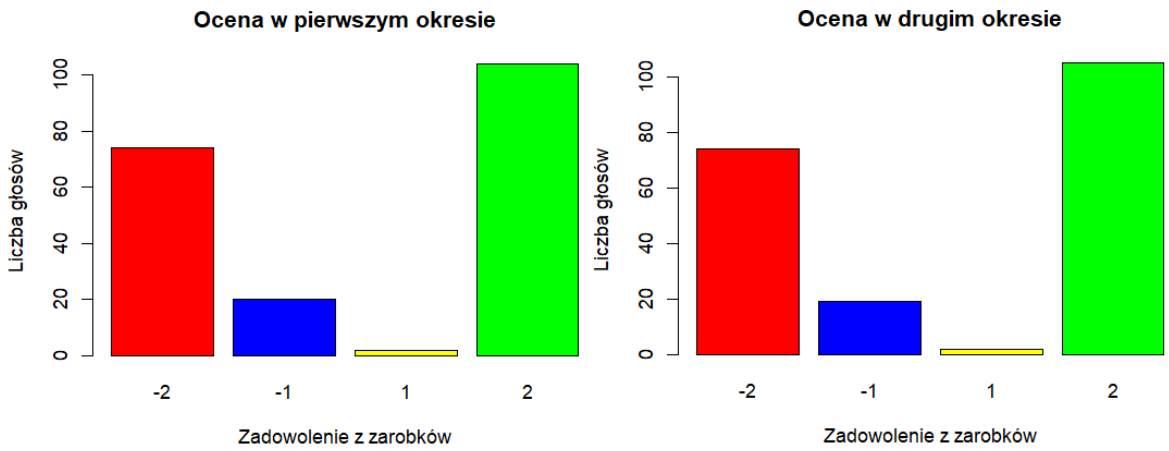


Rysunek 2: Wykresy kołowe dla zmiennej W1 i W2.

Na rysunku 2 możemy zauważyć, że oba wykresy są niezwykle podobne. Świadczy to o podobnym stosunku procentowym głosów oddanych na poszczególne oceny zadowolenia z zarobków w pierwszym oraz drugim okresie badania. Niekoniecznie natomiast świadczyć może o podobnej ilości oddanych głosów.

Następnie przejdziemy do utworzenia wykresów słupkowych.

```
barplot(table(dane$W1), main="Ocena w pierwszym okresie",
xlab="Zadowolenie z zarobków",ylab="Liczba głosów",
col=c("red", "blue", "yellow", "green"))
barplot(table(dane$W2),main="Ocena w drugim okresie",
xlab="Zadowolenie z zarobków",ylab="Liczba głosów",
col=c("red", "blue", "yellow", "green"))
```



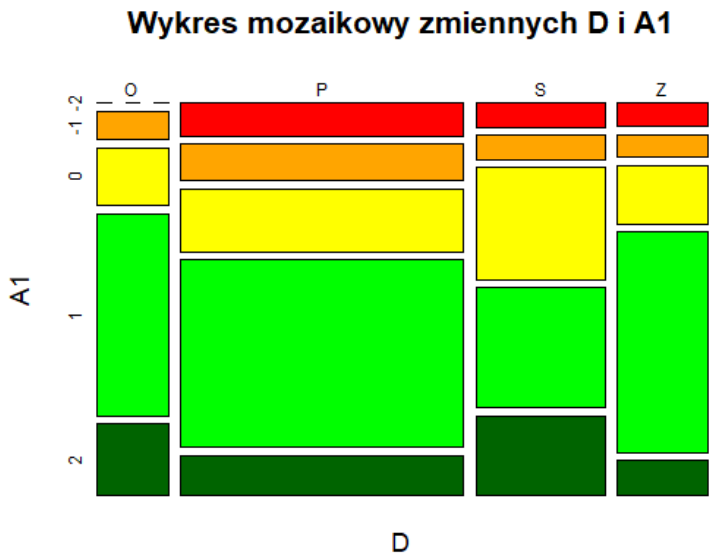
Rysunek 3: Wykresy słupkowe dla zmiennej W1 i W2.

Oczywiście z wykresów słupkowych na rysunku 3 wyciągamy podobne wnioski odnośnie podobieństwa, natomiast wykresy słupkowe dostarczają nam więcej informacji ze względu na oś y, która mówi nam o dokładnych ilościach oddanych głosów. Możemy zatem wnioskować, że nie tylko procentowy stosunek oddanych głosów był podobny, ale również ich ilość.

3.4 Zadanie 4

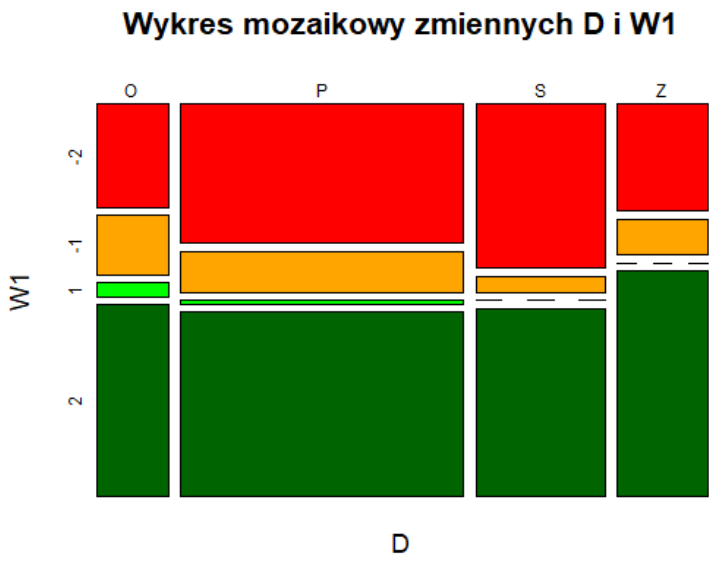
Przy pomocy biblioteki vcd utworzymy odpowiednie wykresy mozaikowe. Wpierw zaczniemy od wykresu dla zmiennych D i A1.

```
mosaicplot(~D+A1,dane,shade=FALSE,color=c('red', 'orange', 'yellow',
'green', 'darkgreen'), main='Wykres mozaikowy zmiennych D i A1')
```



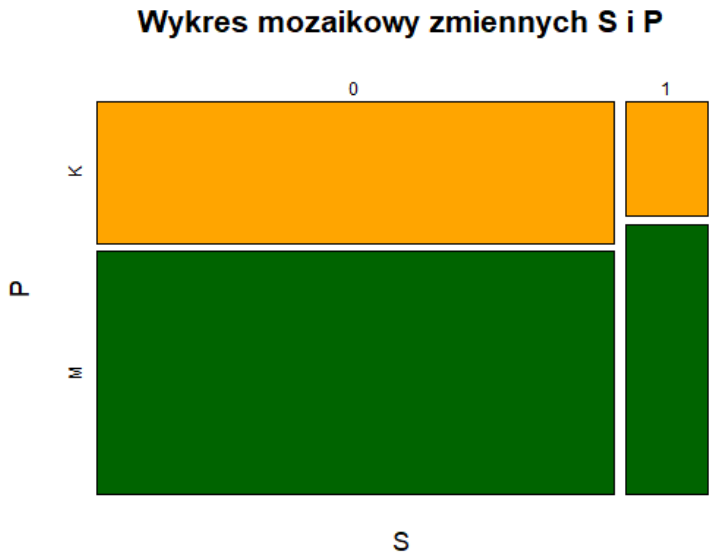
Rysunek 4: Wykres mozaikowy D i A1.

Z rysunku 4, przedstawiającego wykres mozaikowy dla D i A1, odczytać możemy, iż działem z największą ilością oddanych głosów będzie dział produkcyjny. Dodatkowo w większości zgodzili się, że w miejscu pracy panuje bardzo dobra atmosfera. Wyraźna różnica względem innych działów widoczna jest na dziale sprzedaży, w którym podobna ilość ankietowanych wybrała odpowiedź "Trudno powiedzieć" oraz odpowiedź "Zgadzam się". Mimo tego jest to dział z największą stosunkową ilością głosów, w których ankietowani zdecydowanie się zgodzili z bardzo dobrą atmosferą panującą w miejscu pracy.



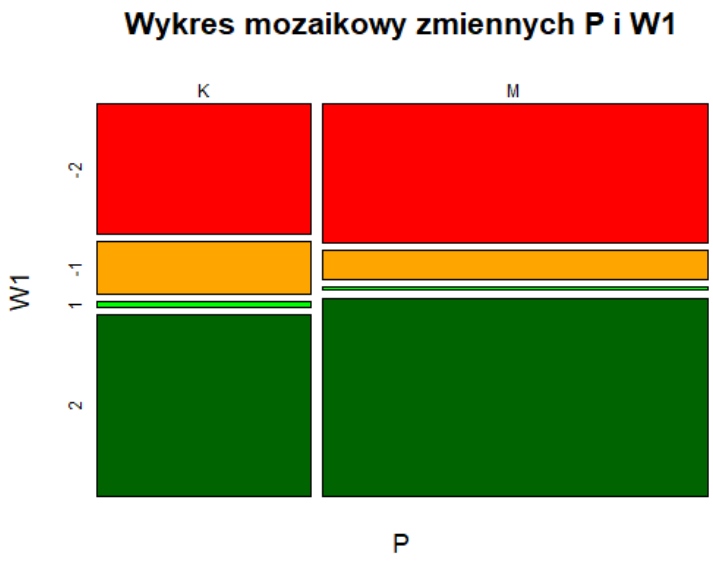
Rysunek 5: Wykres mozaikowy D i W1.

Wykres mozaikowy dla zmiennej D oraz W1 wyraźnie wskazuje, że pracownicy wszystkich działów przeważnie odpowiadają na pytanie odnośnie zadowolenia ze swojego wynagrodzenia "Zdecydowanie się zgadzam" oraz "Zdecydowanie się nie zgadzam". Grupa głosująca bardziej neutralnie stanowi zdecydowaną mniejszość.



Rysunek 6: Wykres mozaikowy S i P.

Z wykresu dla zmiennej S i P odczytać możemy, że większość stanowisk kierowniczych w firmie obejmują mężczyźni. Dodatkowo byli oni grupą częściej udzielającą się w ankiecie.



Rysunek 7: Wykres mozaikowy P i W1.

Na rysunku 7, przedstawiającym wykres zmiennej P i W1, zauważyć możemy, że mężczyźni oraz kobiety odpowiadali procentowo bardzo podobnie na pytanie odnośnie ich zadowolenia z zarobków w pierwszym okresie badania.

4 Część 2

4.1 Zadanie 5

Następną część naszego raportu stanowi losowanie próbki rozmiaru około 1/10 liczby rekordów pewnej bazy danych ze zwracaniem i bez niego. Pomocna tutaj okazuje się funkcja *sample* z pakietu *stats*.

```
sample(x, size, replace = FALSE, prob = NULL)
```

Przyjmuje ona 4 argumenty: bazę danych x, z której nastąpi losowanie; rozmiar próbki *size*, który musi być liczbą całkowitą; informację o tym, czy losowanie jest ze zwracaniem (*replace* = TRUE) lub nie (*replace* = FALSE) oraz wektor wag prawdopodobieństwa poszczególnych elementów bazy *prob*, który pomijamy ze względu na takie same szanse wylosowania poszczególnych elementów.

Dla bazy danych rozpatrywanej w części pierwszej program będzie wyglądał następująco dla losowania ze zwracaniem:

```
zwr1 <- sample(nrow(dane), 0.1*nrow(dane), replace = TRUE),
```

natomiast dla przypadku bez zwracania jak poniżej:

```
zwr0 <- sample(nrow(dane), 0.1*nrow(dane), replace = FALSE).
```

4.2 Zadanie 6

Naszym kolejnym celem jest analiza oceny atmosfery pracy w badanej grupie, a także uwzględniając podział na dział firmy i płeć pracownika. W tym celu zastosowaliśmy funkcje: *summary* oraz *likert.density.plot* i *likert.bar.plot* z biblioteki *likert*.

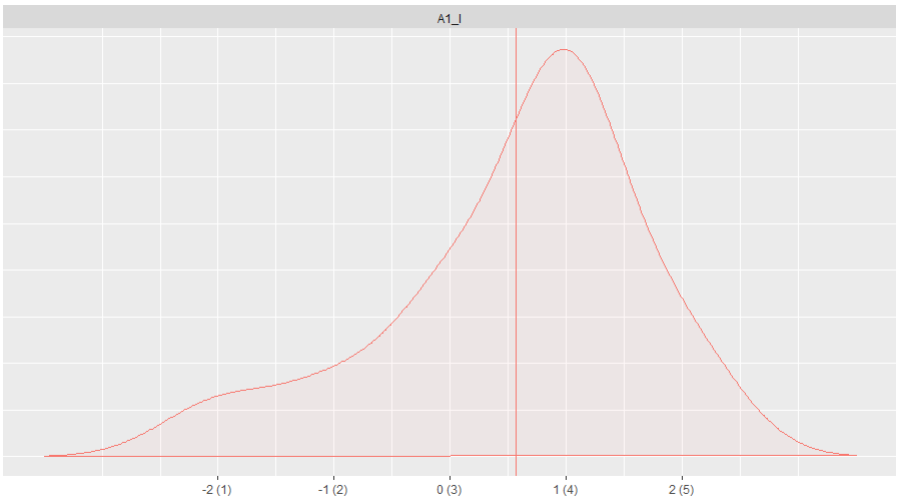
```
X <- likert(dane["A1_l"])

summary(X)
Item  low neutral high  mean       sd
A1_l 15.5      20 64.5 3.565 1.063688

Y <- likert(dane["A2_l"])

summary(Y)
Item  low neutral high mean       sd
A2_l 15.5    17.5   67  3.6 1.041838
```

Rysunek 8: Wyniki funkcji *summary()* dla zmiennych $A1$ i $A2$.

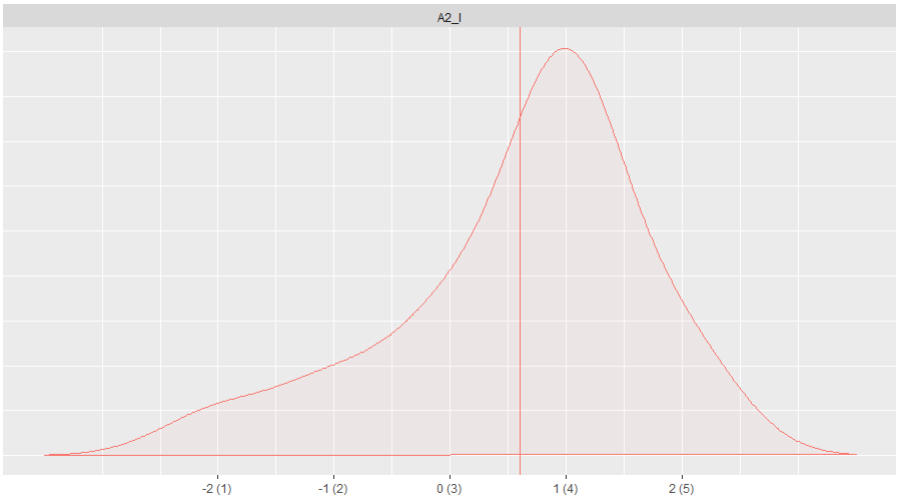
Na powyższym rysunku przedstawiliśmy, za pomocą funkcji *summary()*, odsetki udzielonych odpowiedzi, które zostały podzielone do trzech kategorii. Każda ocena otrzymała swoją rangę, która, w tym przypadku, jest daną oceną powiększoną o 3. Oznacza to, że wektor odpowiedzi $(-2, -1, 0, 1, 2)$ odpowiada wektorowi rang $(1, 2, 3, 4, 5)$. Stąd rangi "1" i "2" zostały zakwalifikowane do oceny "low", "3" oznacza "neutral", natomiast "4" i "5" przydzielono do klasy "high". Ponadto funkcja *summary* obliczyła średnią rangę z uzyskanych odpowiedzi oraz ich odchylenie standardowe.

Analizując wyniki możemy wywnioskować, że większość ankietowanych uważała atmosferę w pracy za bardzo dobrą zarówno za pierwszym, jak i za drugim razem. Porównując oba zestawienia możemy zauważyć, że rok po przeprowadzeniu pierwszej ankiety odsetek osób neutralnie oceniających atmosferę w pracy zmalał o $2,5$ p. proc. przy takim samym wzroście liczby osób pozytywnie oceniających atmosferę. Liczba osób mających odmienne zdanie nie zmieniła się w badanym okresie czasu. To wszystko przyczyniło się do niewielkiego wzrostu średniej rangi odpowiedzi oraz niewielkiego spadku odchylenia standardowego.

Powyższe rozważania zilustrowaliśmy poniżej, rysując wykresy gęstości rozkładu oraz wykresy pudełkowe dla zmiennych $A1$ i $A2$ za pomocą odpowiednio funkcji *likert.density.plot* i *likert.bar.plot*. Widzimy, porównując wykresy gęstości, że oba są jednomodalne, prawostronnie skośne, a ich kształty są bardzo podobne do siebie, aczkolwiek rozkład zmiennej $A2$ jest nieco bardziej leptokurtyczny niż jego odpowiednik dla zmiennej $A1$.

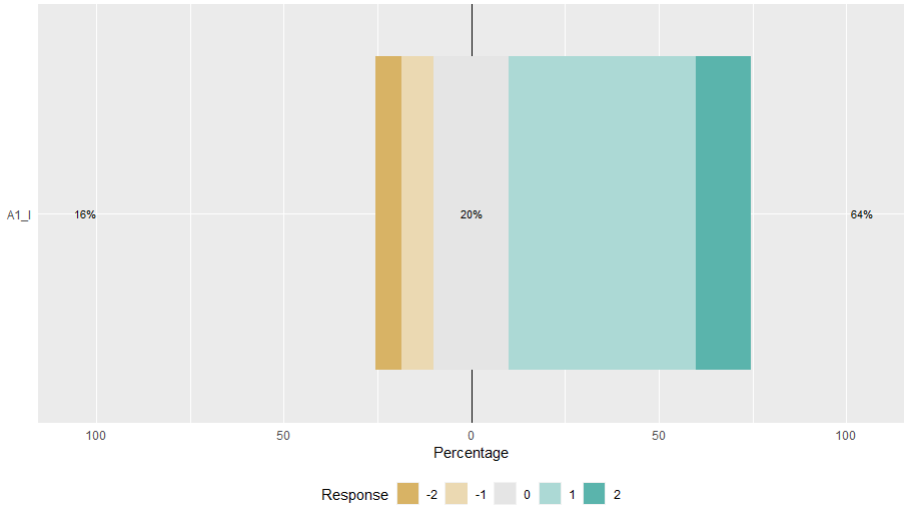


Rysunek 9: Wykres gęstości rozkładu dla zmiennej $A1$.

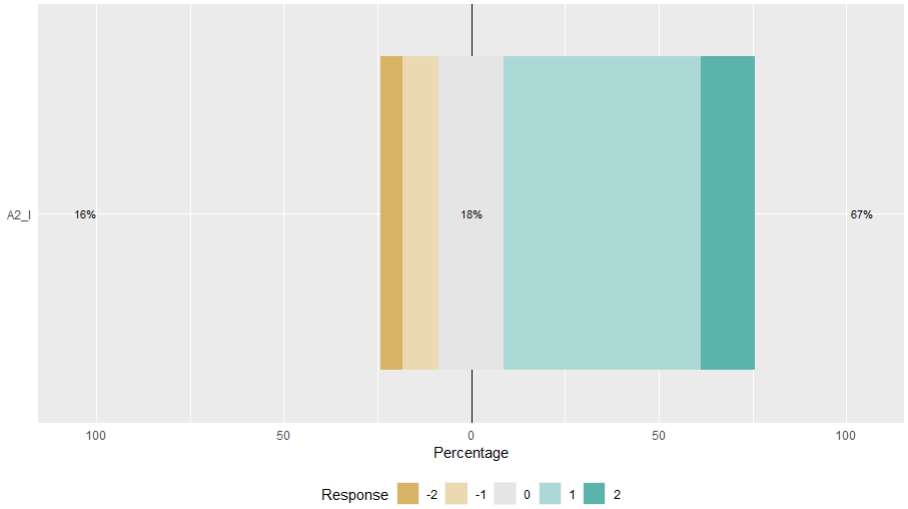


Rysunek 10: Wykres gęstości rozkładu dla zmiennej $A2$.

Podobnie wygląda sytuacja na poniższych wykresach pudełkowych. Na każdym z nich znajdują się trzy liczby: są one niczym innym jak zaokrąglonymi odsetkami odpowiedzi wyświetlonymi powyżej przy pomocy funkcji *summary*.

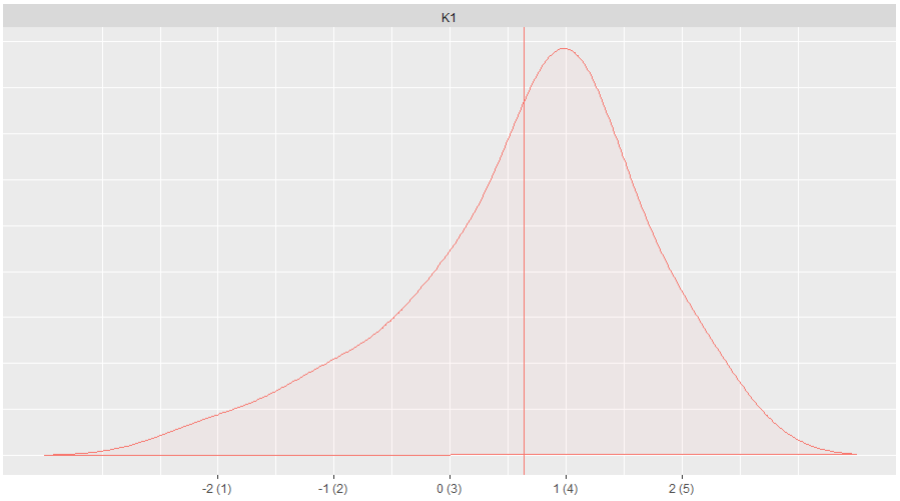


Rysunek 11: Wykres pudełkowy dla zmiennej $A1$.

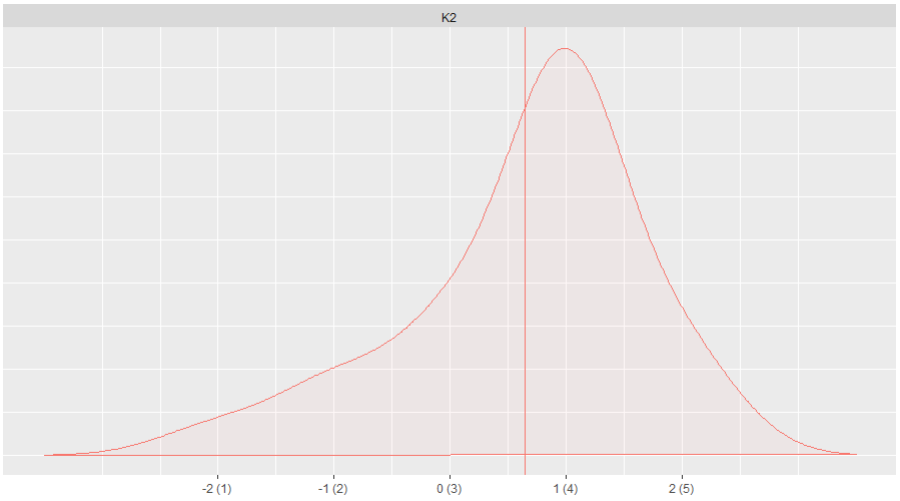


Rysunek 12: Wykres pudełkowy dla zmiennej $A2$.

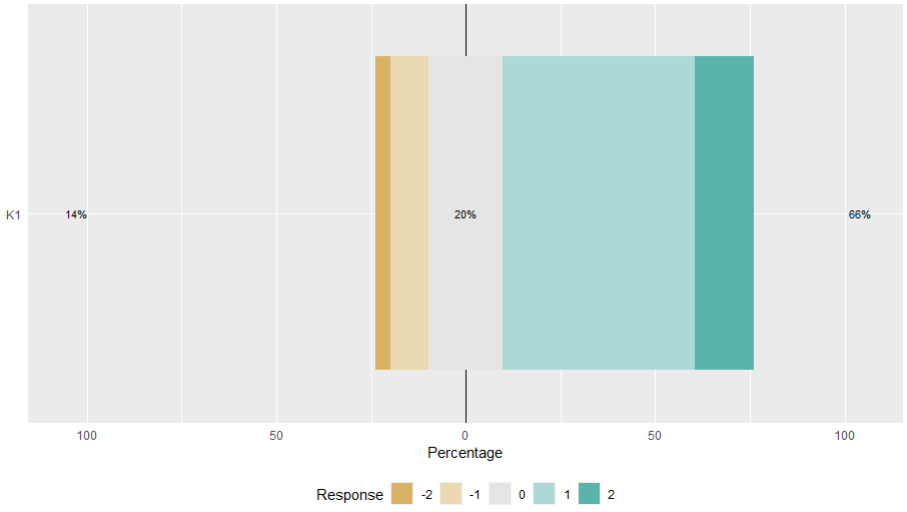
Analizy oceny atmosfery w pracy przeprowadziliśmy również ze względu na płeć pracownika. Wykresy gęstości rozkładu zarówno dla kobiet, jak i mężczyzn w obu porach przeprowadzenia badania, wyglądają podobnie do gęstości dla całej próby. Podobnie jak dla całej populacji, w obu przypadkach po roku odnotowano niewielki wzrost średniego wyniku oraz mały spadek odchylenia standardowego. Ponadto, porównując wykresy pudełkowe z dwóch badań u kobiet i mężczyzn, widzimy ich znaczące podobieństwo. Aczkolwiek, w obu okresach większy odsetek osób niezadowolonych z panującej atmosfery był wśród mężczyzn, a wyższy odsetek osób zadowolonych panował u kobiet.



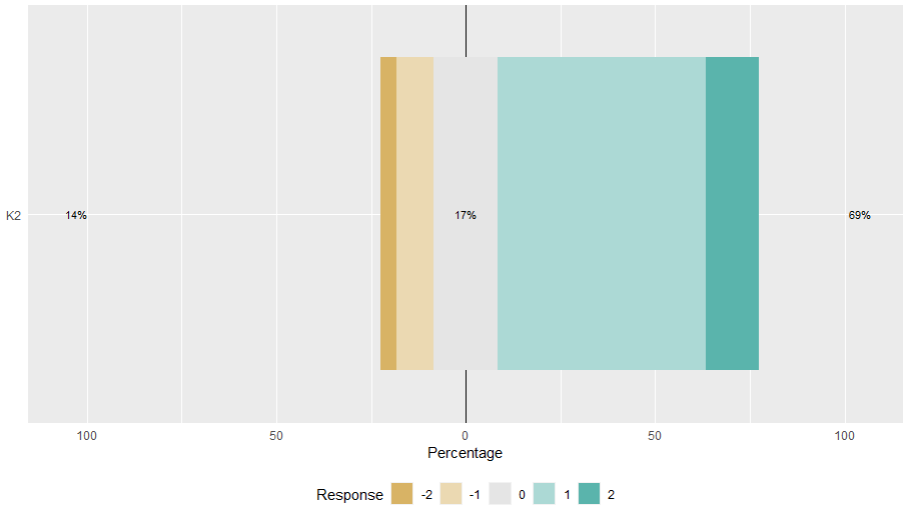
Rysunek 13: Wykres gęstości rozkładu dla zmiennej $A1$ wśród kobiet.



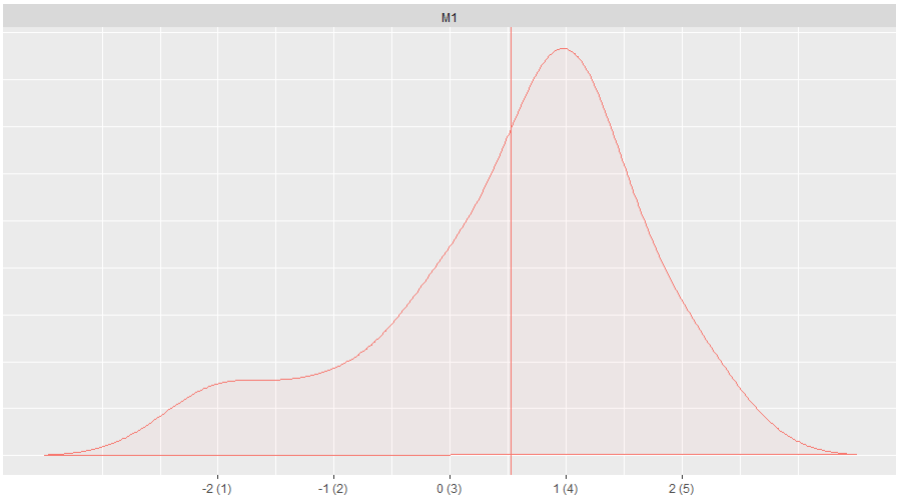
Rysunek 14: Wykres gęstości rozkładu dla zmiennej $A2$ wśród kobiet.



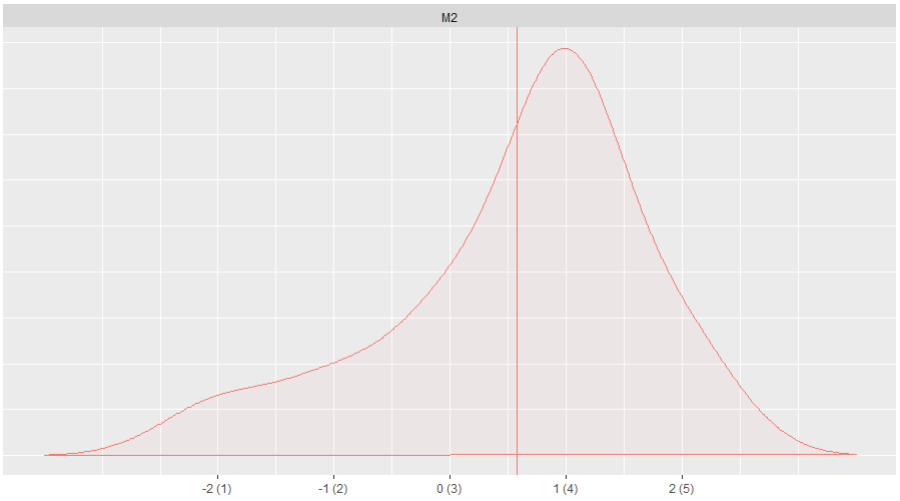
Rysunek 15: Wykres pudełkowy dla zmiennej $A1$ wśród kobiet.



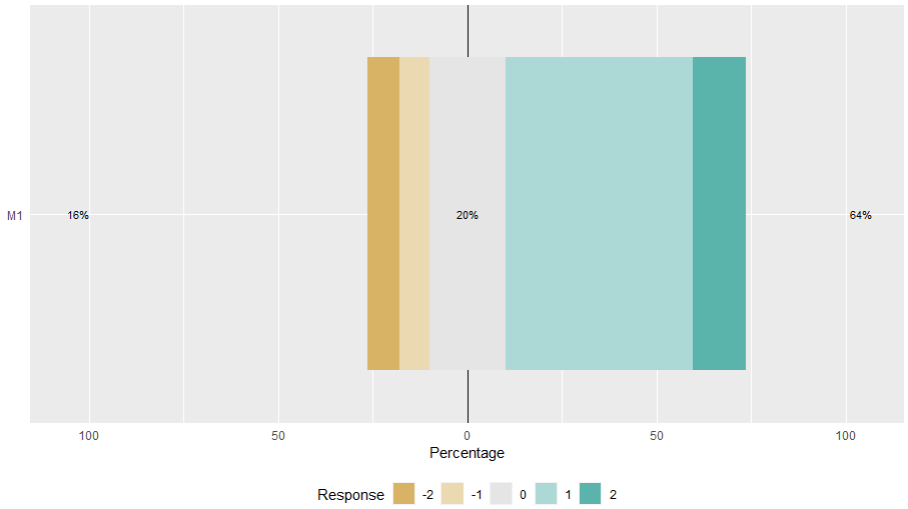
Rysunek 16: Wykres pudełkowy dla zmiennej $A2$ wśród kobiet.



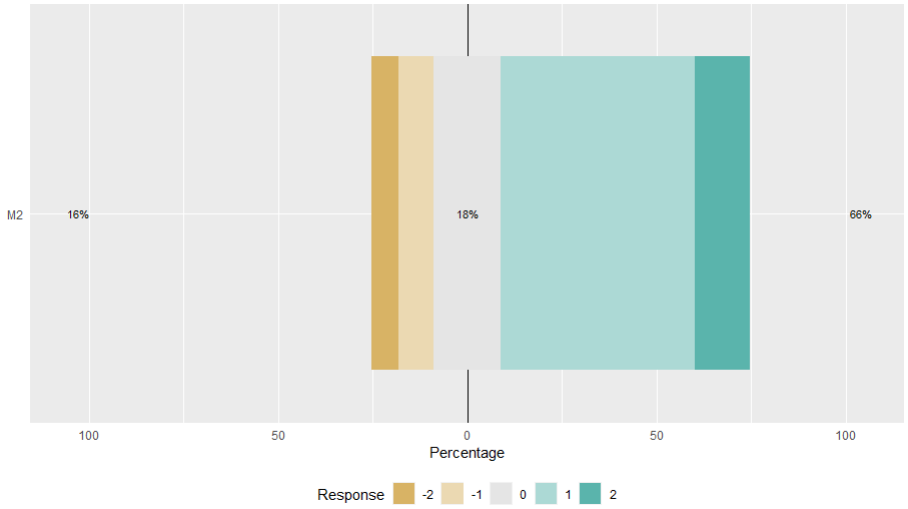
Rysunek 17: Wykres gęstości rozkładu dla zmiennej $A1$ wśród mężczyzn.



Rysunek 18: Wykres gęstości rozkładu dla zmiennej $A2$ wśród mężczyzn.



Rysunek 19: Wykres pudełkowy dla zmiennej $A1$ wśród mężczyzn.

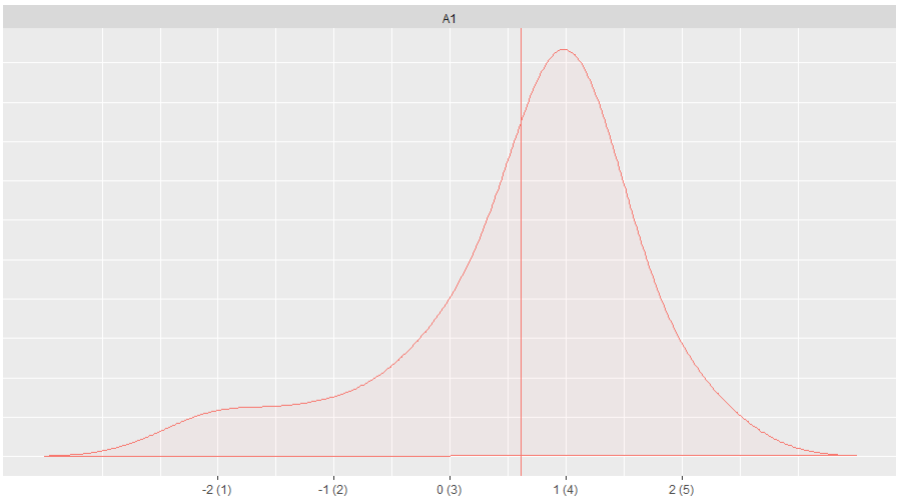


Rysunek 20: Wykres pudełkowy dla zmiennej $A2$ wśród mężczyzn.

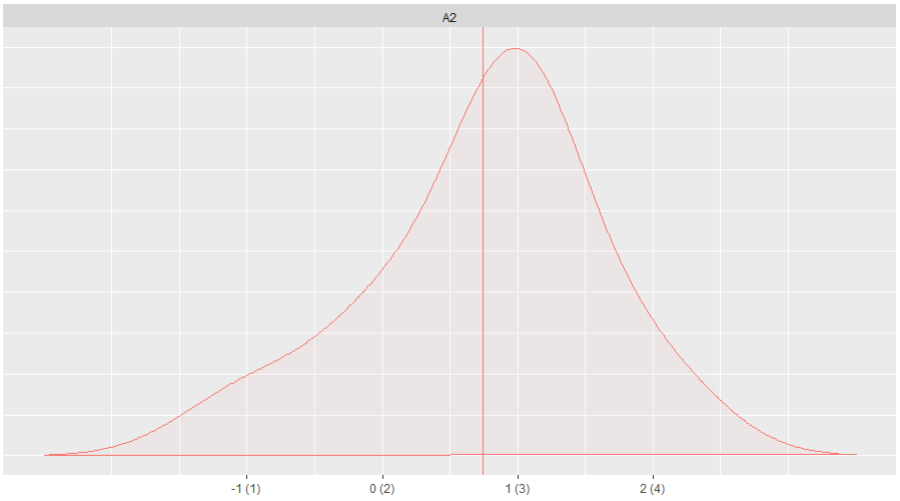
Postanowiliśmy również przeprowadzić analizę danych ze względu na dział, w którym znajdował się dany ankietowany. Dla każdego z nich opisaliśmy wyniki badania, zmiany zaszłe w ciągu roku między turami ankiety oraz narysowaliśmy wykresy gęstości rozkładu oraz pudełkowe.

- Dział zaopatrzenia (Z):

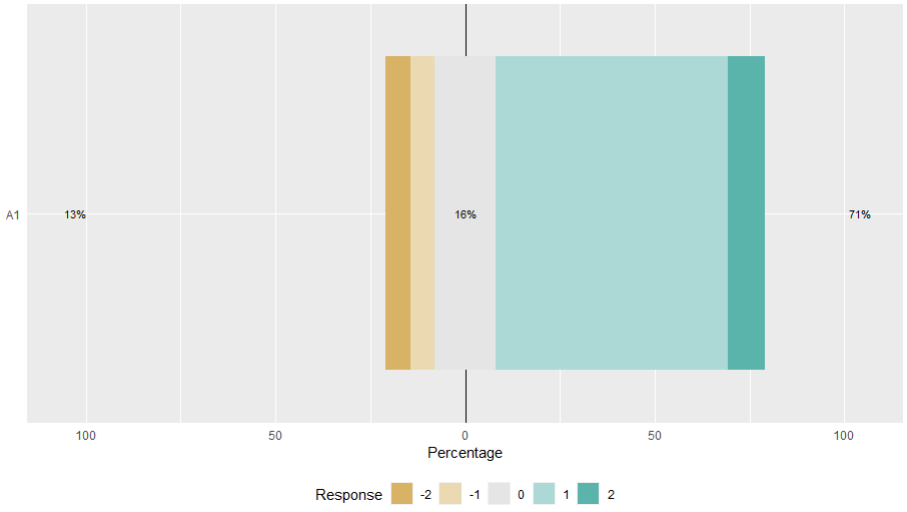
  Podczas pierwszego okresu badania pracownicy działu zaopatrzenia zdecydowanie w większości ocenili atmosferę w pracy na dobrą bądź bardzo dobrą, głosy te stanowiły 71% wszystkich głosów. Za głosem wstrzymało się więcej osób bo 16% , natomiast przeciwko było 13% oddanych głosów. W drugim okresie badania procentowo nie zmieniła się liczba opowiadających się za dobrą atmosferą, jednakże część pracowników zmieniła nastawienie na neutralne z opinii negatywnej, mianowicie 3% łącznych głosujących. Dodatkowo w drugim badaniu żaden z pracowników nie zagłosował na opinię bardzo negatywną. Wykres gęstości w obu okresach się różni, ponieważ lewy ogon rozkładu dla drugiego badania jest zdecydowanie grubszy.



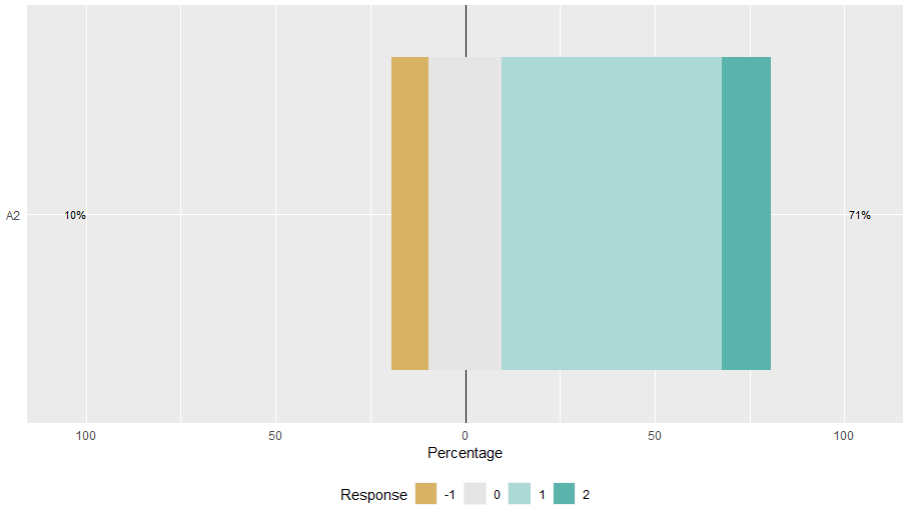
Rysunek 21: Wykres gęstości dla zmiennej $A1$ w dziale zaopatrzenia.



Rysunek 22: Wykres gęstości dla zmiennej $A2$ w dziale zaopatrzenia.



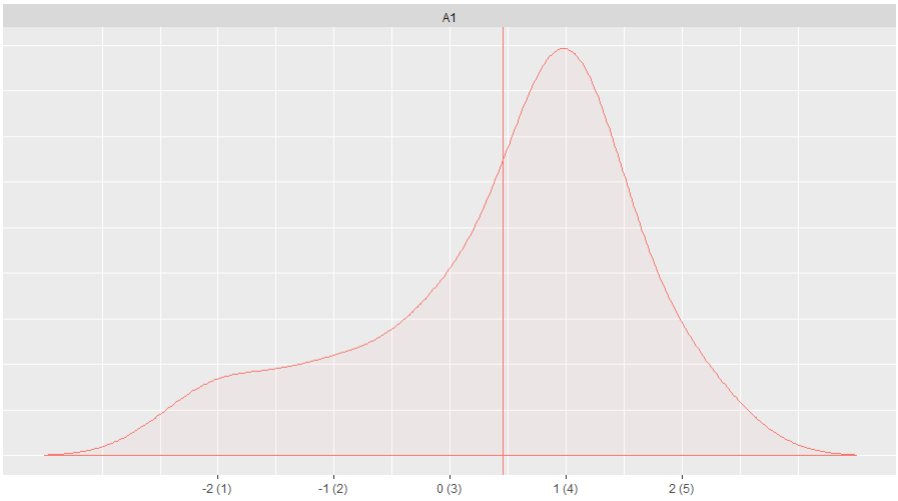
Rysunek 23: Wykres pudełkowy dla zmiennej $A1$ w dziale zaopatrzenia.



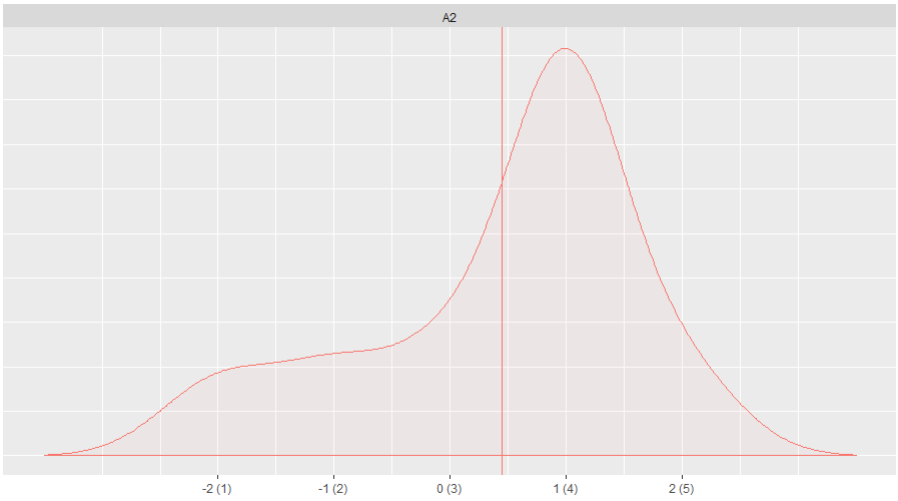
Rysunek 24: Wykres pudełkowy dla zmiennej $A2$ w dziale zaopatrzenia.

- Dział produkcyjny (P):

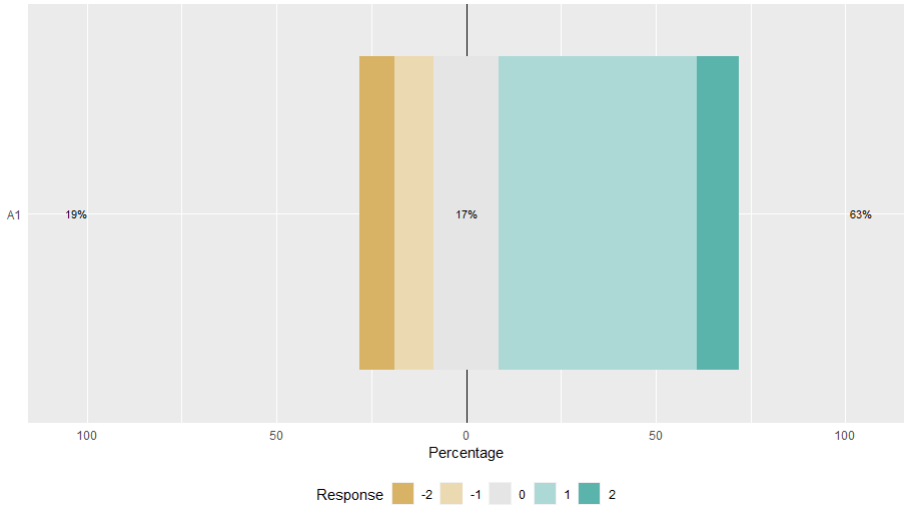
  W dziale produkcji, w pierwszym badaniu, ok. 1/5 badanych nie uznawała atmosfery w pracy za dobrą, podczas gdy ok. 63 % było odmiennego zdania, a pozostałe 17 % nie miało zdania na ten temat. Rozkłady obu zmiennych są podobne do siebie, aczkolwiek wartość średnia w drugim badaniu nieznacznie spadła przy minimalnym wzroście rozproszenia obserwacji. Wpływ na to miał spadek liczby osób z neutralnym poglądem na rzecz wzrostu liczności grup zarówno mających negatywną opinię, jak i pozytywną. Minimalnie więcej osób zmieniło jednak zdanie na negatywne, stąd drobny spadek wartości średniej.



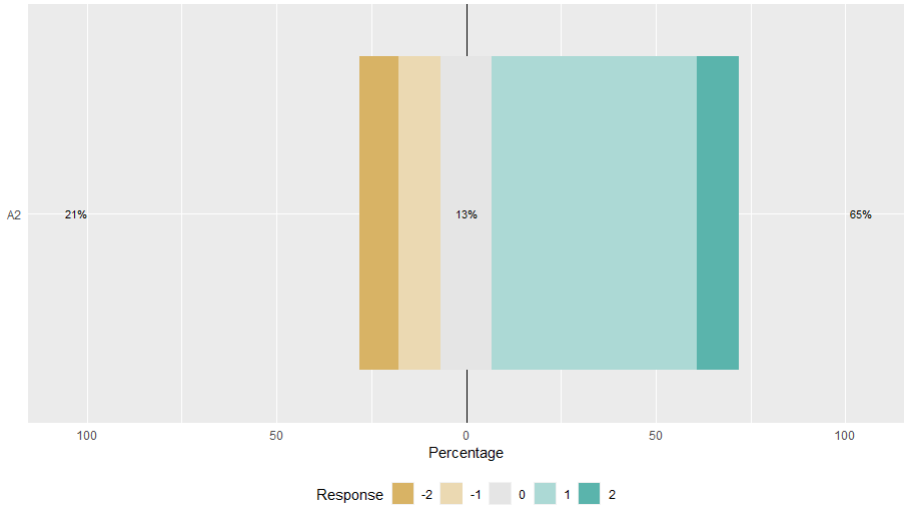
Rysunek 25: Wykres gęstości dla zmiennej $A1$ w dziale produkcyjnym.



Rysunek 26: Wykres gęstości dla zmiennej $A2$ w dziale produkcyjnym.



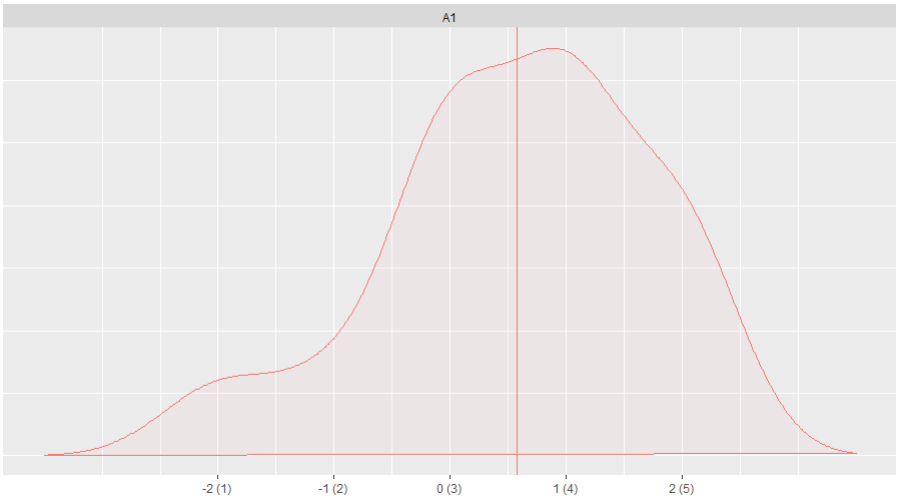
Rysunek 27: Wykres pudełkowy dla zmiennej $A1$ w dziale produkcyjnym.



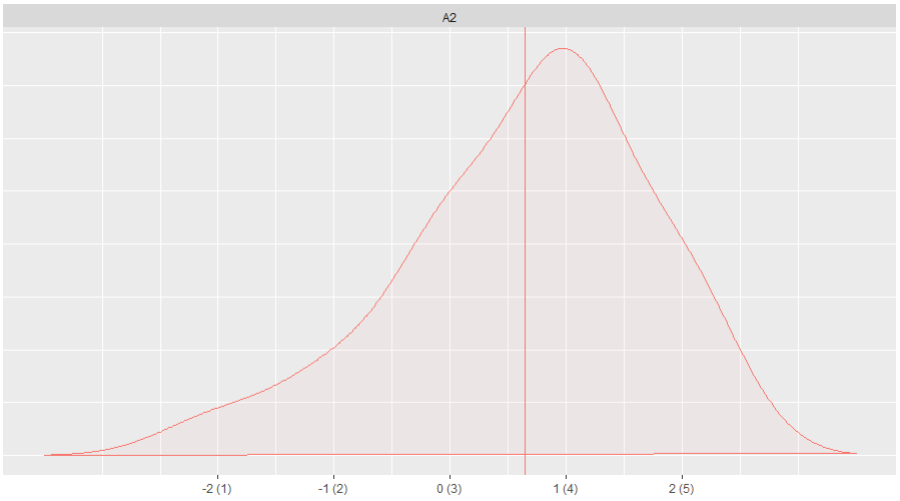
Rysunek 28: Wykres pudełkowy dla zmiennej $A2$ w dziale produkcyjnym.

- Dział sprzedaży (S):

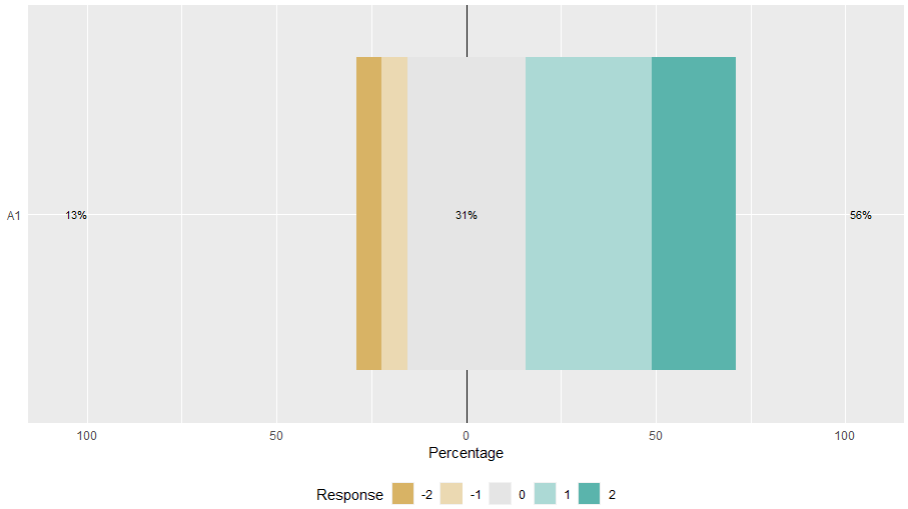
  Nieco inaczej wygląda sytuacja w dziale sprzedaży. W obu badaniach ok. 13 % ankietowanych nie zgodziło się ze zdaniem nt. dobrej atmosfery w pracy. 56 % osób zgodziło się z tą opinią, a pozostałe 31 % zajęło neutralne stanowisko. W przeciągu roku odsetek osób niezgadzających się z tym stwierdzeniem nie zmienił się, podczas gdy odsetek pracowników uznających atmosferę za dobrą zwiększył się kosztem grupy „neutralnej" o 6.7 p. proc. Rozkłady zmiennych $A1$ i $A2$ różnią się od siebie. Ogony pierwszego rozkładu są cięższe od ich odpowiedników przy zmiennej $A2$, natomiast drugi rozkład przypomina kształtem rozkład normalny. Ponadto, rozkład ocen atmosfery pracy w drugim badaniu charakteryzuje się wyższą średnią i wyższym skupieniem odpowiedzi (mniejsze odchylenie standardowe).



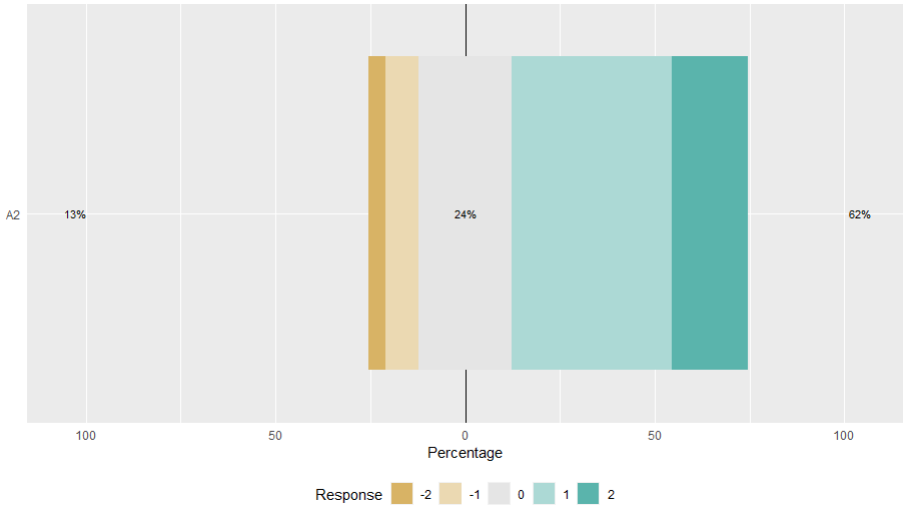
Rysunek 29: Wykres gęstości dla zmiennej $A1$ w dziale sprzedaży.



Rysunek 30: Wykres gęstości dla zmiennej $A2$ w dziale sprzedaży.



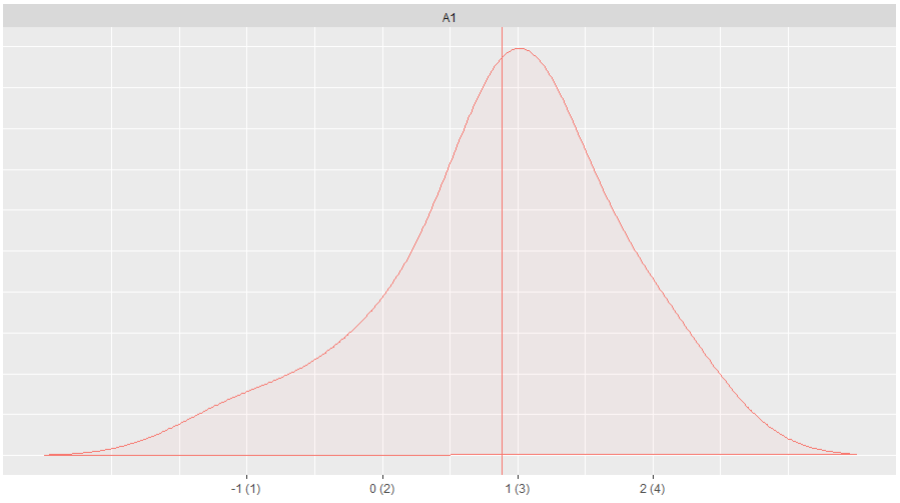
Rysunek 31: Wykres pudełkowy dla zmiennej $A1$ w dziale sprzedaży.



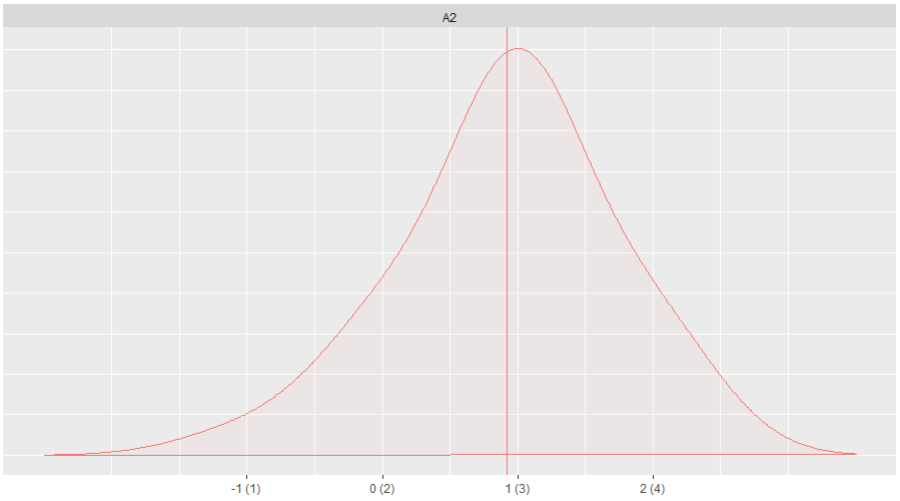
Rysunek 32: Wykres pudełkowy dla zmiennej $A2$ w dziale sprzedaży.

- Dział obsługi kadrowo-płacowej (O):

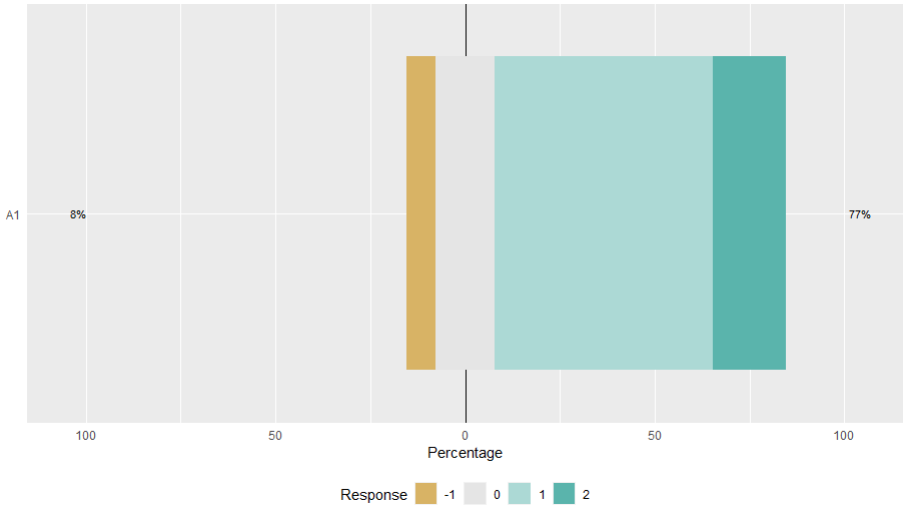
  Dział obsługi kadrowo-płacowej charakteryzuje się natomiast tym, że średnia głosów w obu okresach badania wynosiła blisko wartość 1. Jest to najwyższa średnia głosów spośród innych działów poddanym ankietowaniu. Procent osób głosujących na odpowiedź pozytywną się nie zmienił. Odpowiedzi negatywne zaliczyły jednak spadek z 8% na 4%, co znaczy również, że większy odsetek ankietowanych wybrało odpowiedź "nie wiem".



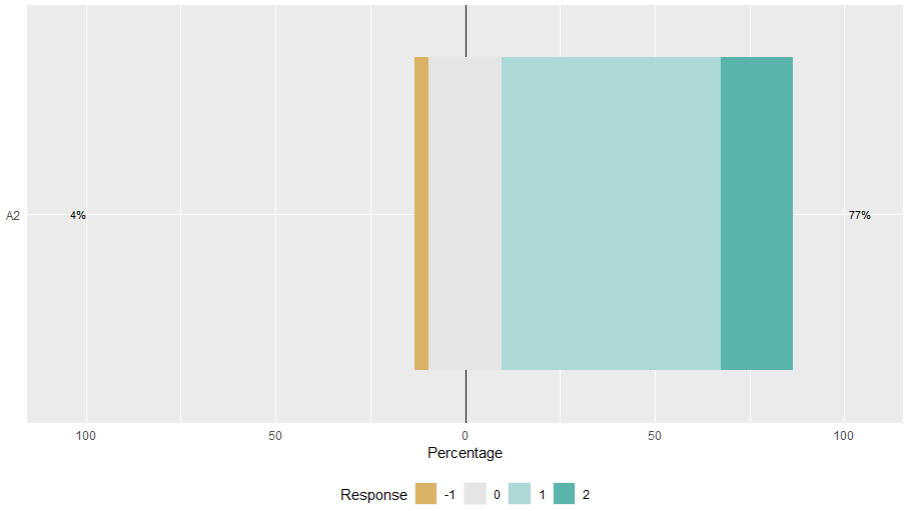
Rysunek 33: Wykres gęstości dla zmiennej $A1$ w dziale obsługi kadrowo-płacowej.



Rysunek 34: Wykres gęstości dla zmiennej $A2$ w dziale obsługi kadrowo-płacowej.



Rysunek 35: Wykres pudełkowy dla zmiennej $A1$ w dziale obsługi kadrowo-płacowej.



Rysunek 36: Wykres pudełkowy dla zmiennej $A2$ w dziale obsługi kadrowo-płacowej.

4.3 Zadanie 7

Kolejnym zadaniem było stworzenie programu wyznaczającego realizację przedziału ufności Cloppera-Pearsona dla pewnego zadanego poziomu ufności. Jego kod został przedstawiony poniżej.

```
  Clopper <- function(alpha, x, n) {
  l <- qbeta(alpha/2, x, n-x+1)
  u <- qbeta(1-alpha/2, x+1, n-x)
  if (x==0) {
    L <- 0
  } else {
    L <- l
  }
  if (x==n) {
    P <- 1
  } else {
    P <- u
  }
  paste("Dolna granica: ", L, "Górna granica: ", P)
}
```

Teraz zastosujemy powyższą funkcję do wyznaczenia przedziałów ufności dla prawdopodobieństwa, że pracownik jest zadowolony ze swojego wynagrodzenia w pierwszym badanym okresie. Powstałe wyniki porównamy z ich odpowiednikami wyznaczonymi za pomocą funkcji *binom.confint* z pakietu *binom*. Te czynności wykonamy też w podgrupach, biorąc pod uwagę dział oraz stanowisko pracownika.

We wszystkich przypadkach x, n i alpha oznaczają kolejno: liczbę sukcesów (w naszym przypadku liczbę osób zadowolonych ze swojego wynagrodzenia), liczność próby oraz poziom istotności, który przyjęliśmy na poziomie $alpha = 0.05$.

- Przypadek 1.: cała grupa

```
 binom.confint(x, n, conf.level=1-alpha, methods="exact")
  method  x  n   mean  lower     upper
1  exact 106 200 0.53 0.4583305 0.6007671

 Clopper(alpha,x,n)
 [1] "Dolna granica:  0.458330500411475
      Górna granica:  0.600767058802886"
```

- Przypadek 2.: podział ze względu na dział:
  - Dział zaopatrzenia:

```
   binom.confint(x, n, conf.level=1-alpha, methods="exact")
  method x n     mean      lower     upper
1  exact 5 31 0.1612903 0.05452433 0.3372716

   Clopper(alpha,x,n)
[1] "Dolna granica:  0.0545243262150835
     Górna granica:  0.337271584973179"
```

  - Dział produkcyjny:

```
   binom.confint(x, n, conf.level=1-alpha, methods="exact")
  method x n    mean      lower     upper
1  exact 4 98 0.04081633 0.01123132 0.1012178
```

```
   Clopper(alpha,x,n)
[1] "Dolna granica:  0.0112313175601852
     Górna granica:  0.101217818888766"
```

– Dział sprzedaży:

```
   binom.confint(x, n, conf.level=1-alpha, methods="exact")
  method x n    mean      lower     upper
1 exact 4 45 0.08888889 0.02475296 0.2122117

   Clopper(alpha,x,n)
[1] "Dolna granica:  0.0247529577834691
     Górna granica:  0.21221173695434"
```

– Dział obsługi kadrowo-płacowej:

```
   binom.confint(x, n, conf.level=1-alpha, methods="exact")
 method x  n   mean      lower     upper
1 exact 5 26 0.1923077 0.06554811 0.3935055

   Clopper(alpha,x,n)
[1] "Dolna granica:  0.0655481087367825
     Górna granica:  0.393505527939322"
```

- Przypadek 3.: podział ze względu na stanowisko:
  – stanowiska kierownicze:

```
   binom.confint(x, n, conf.level=1-alpha, methods="exact")
  method x  n    mean      lower     upper
1  exact 6 27 0.2222222 0.08621694 0.4225831

   Clopper(alpha,x,n)
[1] "Dolna granica:  0.0862169399475668
     Górna granica:  0.422583060044379"
```

– pozostałe stanowiska:

```
   binom.confint(x, n, conf.level=1-alpha, methods="exact")
  method x  n     mean       lower      upper
1  exact 3 173 0.01734104 0.003590516 0.04983927

   Clopper(alpha,x,n)
[1] "Dolna granica:  0.00359051636947693
     Górna granica:  0.0498392679898603"
```

Przyrównując do siebie wyniki funkcji *Clopper* i *binom.confint* możemy stwierdzić ich identyczność.

5 Część trzecia

5.1 Zadanie 8

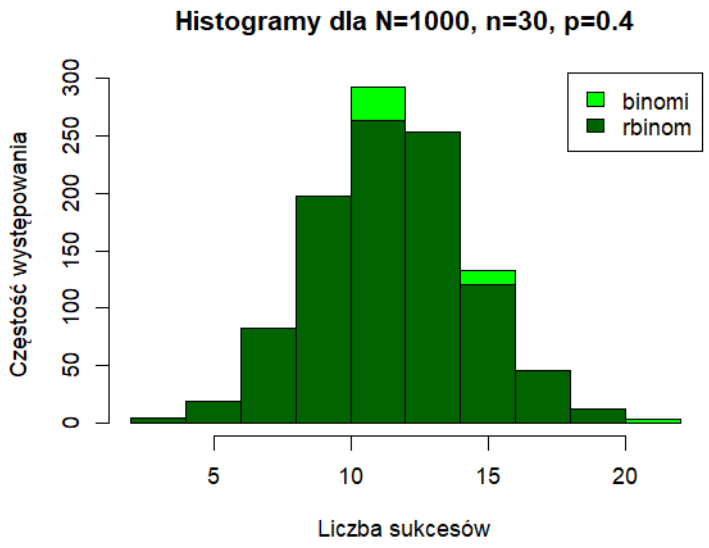
W zadaniu ósmym musimy zaproponować algorytm generowania liczb z rozkładu dwumianowego oraz udowodnić jego poprawność. Do generowania liczb zaproponujemy następujący algorytm, gdzie n oznacza liczbę prób, a p prawdopodobieństwo sukcesu w jednej próbie.

1. Generuj wektor $U = (U_1, ..., U_n)$, gdzie zmienne losowe są iid, $U_i \sim U(0, 1)$
2. Sumujemy indykatory zdarzeń $\{U_i < p\}$
3. Powtarzamy N razy

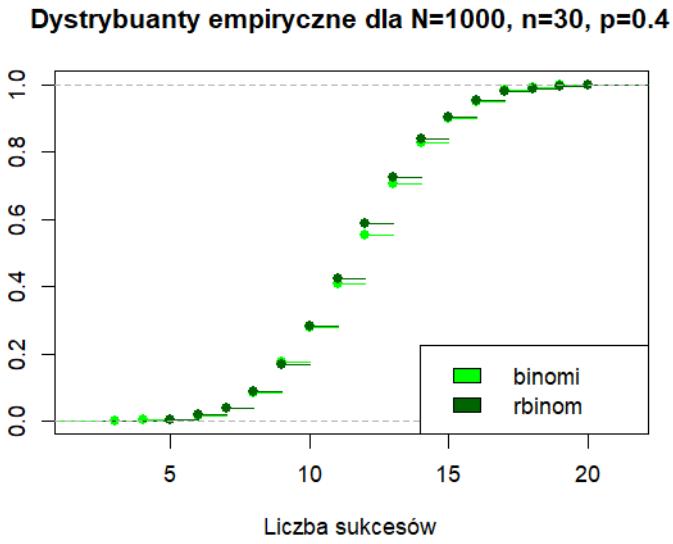
Powyższy zaprogramowany algorytm w języku R.

```
binomi<-function(N,n,p){
  bin<-c()
  for (i in 1:N){
    U<-runif(n,min=0,max=1)
    inds<-sum(U<p)
    bin<-append(bin,inds)
  }
  return(bin)
}
```

W celu sprawdzenia poprawności napisanego przez nas algorytmu porównamy go z funkcją wbudowaną z pakietu R, mianowicie rbinom, która generuje liczby z rozkładu dwumianowego dla zadanych argumentów. Porównamy wynik wywołania obu funkcji na histogramach, jak i również spojrzymy na dystrybuanty empiryczne otrzymane w wyniku obu symulacji. Wpierw obie funkcje przetestujemy dla $N = 1000$ wywołań Monte Carlo oraz dla parametru długości próby $n = 30$ i parametru prawdopodobieństwa sukcesu $p = 0.4$.

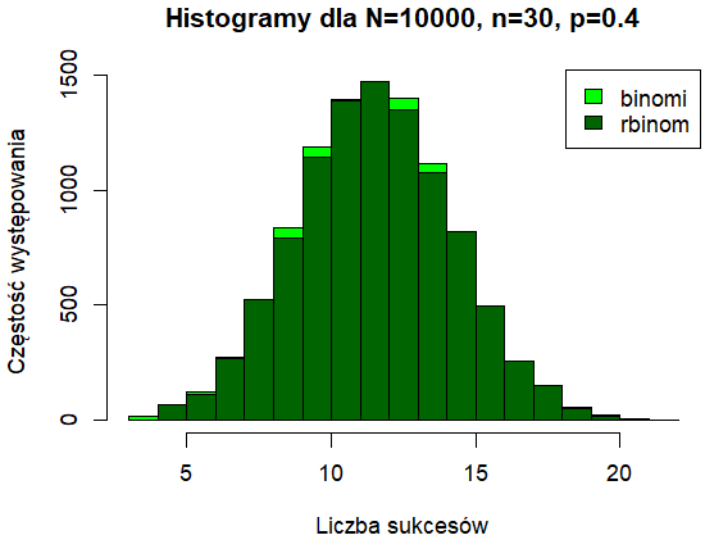


Rysunek 37: Histogramy obu funkcji dla 1000 powtórzeń Monte Carlo.

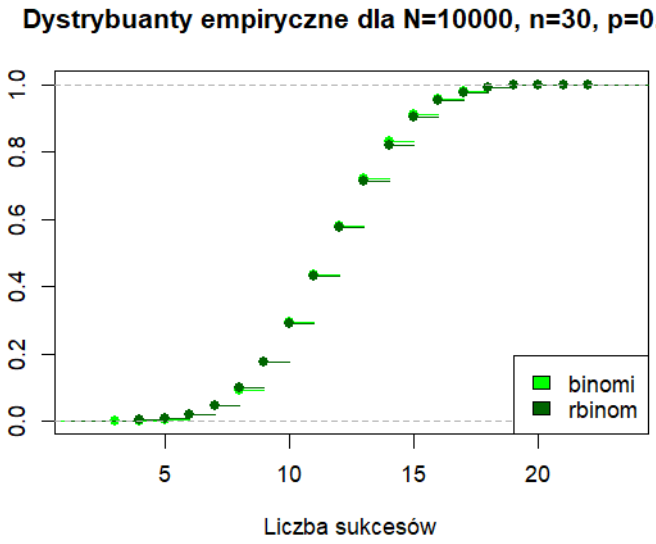


Rysunek 38: Dystrybuanty empiryczne obu funkcji dla 1000 powtórzeń Monte Carlo.

Na wyżej przedstawionym histogramie oraz wykresie dystrybuant (rysunki 37 i 38) zauważamy wyraźne podobieństwo między generatorem liczb z pakietu R oraz autorskim. Możemy zatem wnioskować, iż napisany algorytm działa poprawnie. Sprawdzimy jednak czy przy większej próbie N różnice te będą jeszcze bardziej zanikać, a wykresy i histogramy będą się całkowicie pokrywać.



Rysunek 39: Histogramy obu funkcji dla 10000 powtórzeń Monte Carlo.



Rysunek 40: Dystrybuanty empiryczne obu funkcji dla 10000 powtórzeń Monte Carlo.

Przy $N = 10000$ powtórzeniach Monte Carlo widzimy, że dystrybuanty, podobnie jak histogramy nachodzą na siebie w bardzo dużym stopniu. Wraz ze wzrostem próby odchyły między wykresami są coraz mniejsze. Wnioskujemy zatem, że nie ma podstawy by uznać skonstruowany algorytm za niepoprawny.

5.2 Zadanie 9

Celem dziewiątego zadania jest przeprowadzenie symulacji, której celem jest porównanie prawdopodobieństwa pokrycia i długości przedziałów ufności Cloppera-Pearsona, Walda i trzeciego dowolnego typu przedziału ufności zaimplementowanego w funkcji binom.confint pakietu binom. Trzecim wybranym przez nas typem jest przedział ufności Agrestiego-Coulla z poprawką na ciągłość.

Korzystając ze wzorów podanych na wykładzie napiszemy wzory na przedziały ufności na poziomie ufności $1-\alpha$.

1. Przedział ufności Cloppera-Pearsona (dokładne) $[p_{-\alpha}(X), \bar{p}_{\alpha}(X)]$, gdzie

$$p_{-\alpha}(X) = \begin{cases} 0, \text{ gdy } X = 0, \\ \text{kwantyl rzędu } \alpha/2 \text{ rozkładu } Be(X, n-X+1), \end{cases}$$

$$\bar{p}_{\alpha}(X) = \begin{cases} 1, \text{ gdy } X = n, \\ \text{kwantyl rzędu } 1-\alpha/2 \text{ rozkładu } Be(X+1, n-X). \end{cases}$$

2. Przedział ufności Walda (asymptotyczne) $[T_L^A, T_U^A]$, gdzie

$$T_L^A = \overline{X} - c_\alpha[\overline{X}(1-\overline{X})]^{1/2},$$
$$T_U^A = \overline{X} + c_\alpha[\overline{X}(1-\overline{X})]^{1/2}.$$

   Przy czym $\overline{X} = \sum_{i=1}^n X_i/n$ to średnia próbkowa, a $c_\alpha = \frac{z_{1-\alpha/2}}{\sqrt{n}}$, gdzie α związane jest z przyjętym poziomem ufności $1-\alpha$, natomiast $z_{1-\alpha/2}$ to kwantyl standardowego rozkładu normalnego na poziomie $1-\alpha/2$.

3. Przedział ufności Agrestiego-Coulla (skorygowane Walda, poprawka na ciągłość) $[T_L^{AC}, T_U^{AC}]$, gdzie

$$T_L^{AC} = \tilde{p} - \kappa(\alpha)(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2},$$
$$T_U^{AC} = \tilde{p} + \kappa(\alpha)(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2}.$$

   Przy czym oznaczenia w skorygowanych przedziałach zostały zmienione na $\kappa(\alpha) = z_{1-\alpha/2}$, $\tilde{X} = \sum_{i=1}^n X_i + \kappa^2(\alpha)/2$, $\tilde{n} = n + \kappa^2(\alpha)$, $\tilde{p} = \tilde{X}/\tilde{n}$ oraz $\tilde{q} = 1 - \tilde{p}$.

Wiedząc już jak wyglądają nasze przedziały możemy je zaimplementować. Skorzystamy natomiast z wbudowanej funkcji w pakiecie R binom.confint. Wybranie metody "exact" zwróci nam przedziały ufności Cloppera-Pearsona, metoda "asymptotic" przedziały Walda, natomiast metoda "ac" przedziały Agrestiego-Coulla.

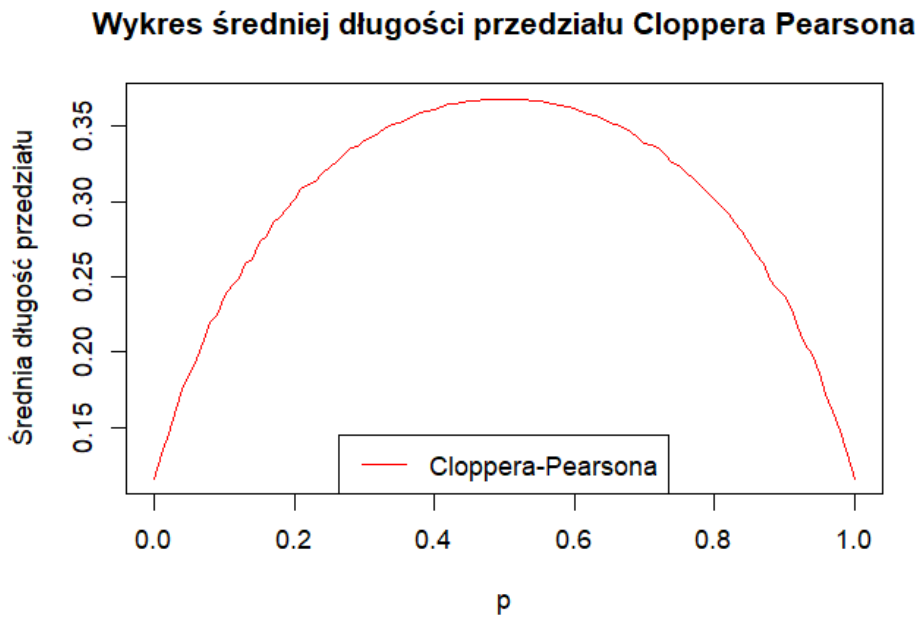
```
binom.confint(x,n,conf.level=0.95,methods="exact")
binom.confint(x,n,conf.level=0.95,methods="asymptotic")
binom.confint(x,n,conf.level=0.95,methods="ac")
```

Aby sprawdzać średnie długości tych przedziałów w zależności od p (prawdopodobieństwa sukcesu) wykorzystamy symulację Monte Carlo. Zaprezentujemy kod odpowiadający za implementację symulacji oraz wyświetlenie wykresu dla przedziałów Cloppera-Pearsona dla rozmiaru próby $n = 30$. Wykorzystamy dodatkowo bilbioteki binom oraz lattice.

```
library(binom)
library(lattice)
srednia_dlugosc <- list()
p_list <- seq(0, 1, by=0.01)
for (i in p_list){
  N <- 1000
  n <- 30
  p <- i
  x <- rbinom(N, n, p)
  result <- binom.confint(x, n, conf.level=0.95, methods="exact")
  len <- result["upper"] - result["lower"]
```

```
  srednia <- sum(len)/N
  srednia_dlugosc <- append(srednia_dlugosc, srednia)
}

plot(p_list, srednia_dlugosc, xlab="p", ylab="Średnia długość przedziału",
     main="Wykres średniej długości przedziału Cloppera Pearsona",
     type = "l", col=('red'))

legend("bottom",legend=c("Cloppera-Pearsona"), col=c('red'), lty=1:1, cex=1)
```

Wywołanie powyższej funkcji zwróci nam następujący wynik.

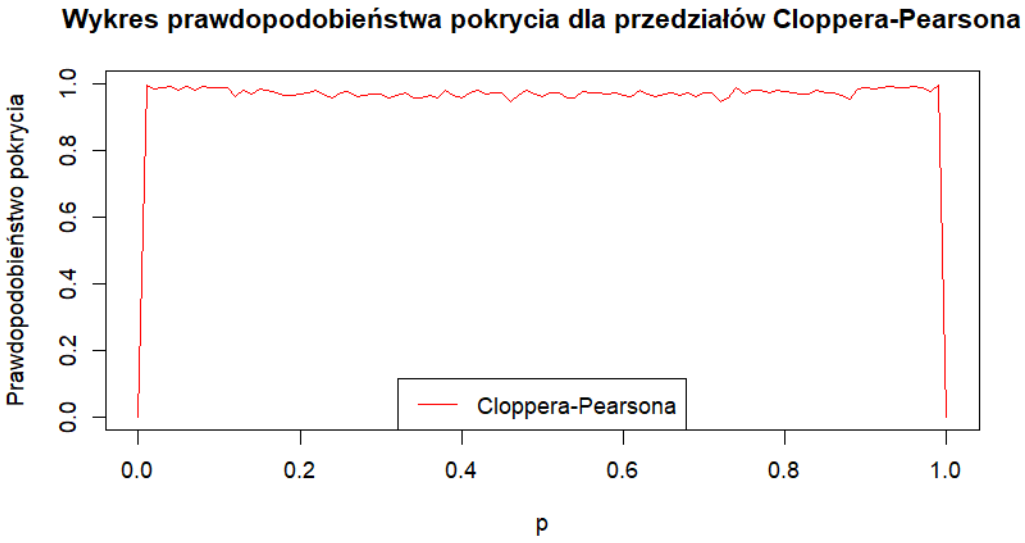


Rysunek 41: Wykres średniej długości przedziału ufności Cloppera-Pearsona w zależności od p dla rozmiaru próby $n = 30$.

Brakuje nam teraz jedynie funkcji odpowiadającej za prawdopodobieństwo pokrycia przedziałów. Zaimplementujemy ją i wyświetlimy odpowiedni wykres przy pomocy następującej funkcji.

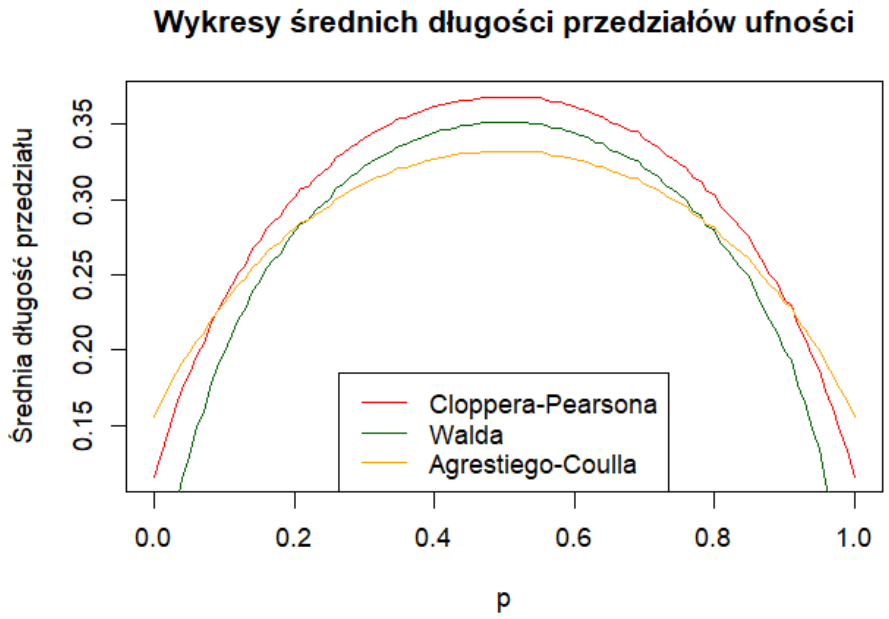
```
pr_pokrycia <- list()
p_list <- seq(0, 1, by=0.01)
for (i in p_list){
  n <- 30
  N <- 1000
  p <- i
  x <- rbinom(N, n, p)
  result <- binom.confint(x, n, conf.level=0.95, methods="exact")
  p_in <- p > result["lower"] & p < result["upper"]
  pokrycie <- sum(p_in)/N
  pr_pokrycia <- append(pr_pokrycia, pokrycie)
}

plot(p_list, pr_pokrycia, xlab="p", ylab="Prawdopodobieństwo pokrycia",
     main="Wykres prawdopodobieństwa pokrycia dla przedziałów Cloppera-Pearsona",
     type = "l", col=('red'))

legend("bottom",legend=c("Cloppera-Pearsona"), col=c('red'), lty=1:1, cex=1)
```

Po wywołaniu funkcji otrzymujemy poniższy wykres.

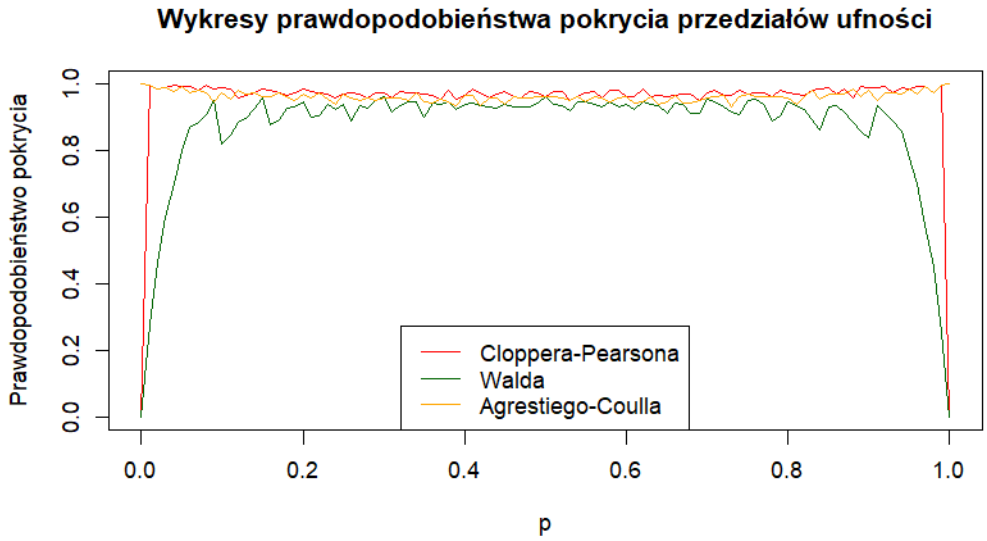


Rysunek 42: Wykresy prawdopodobieństwa pokrycia przedziałów ufności w zależności od p dla rozmiaru próby $n = 30$.

Wiedząc jak wyznaczać średnie długości przedziałów oraz ich prawdopodobieństwo pokrycia porównamy ze sobą przedziały Cloppera-Pearsona, Walda oraz Agrestiego-Coulla dla różnych rozmiarów prób. Wykorzystamy poprzednio napisane funkcje zmieniając jedynie odpowiednio metody.



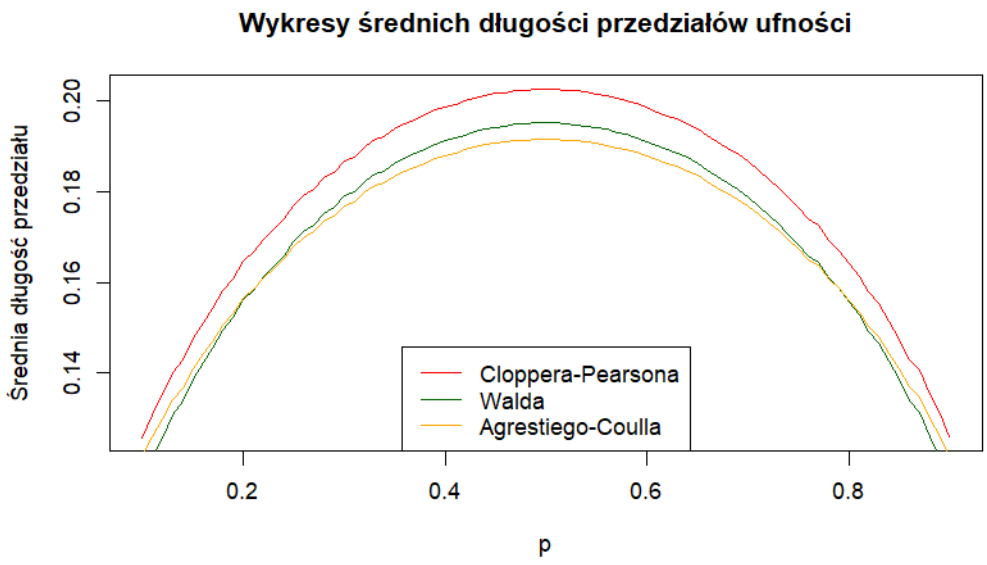
Rysunek 43: Wykresy średniej długości przedziałów ufności w zależności od p dla rozmiaru próby $n = 30$.



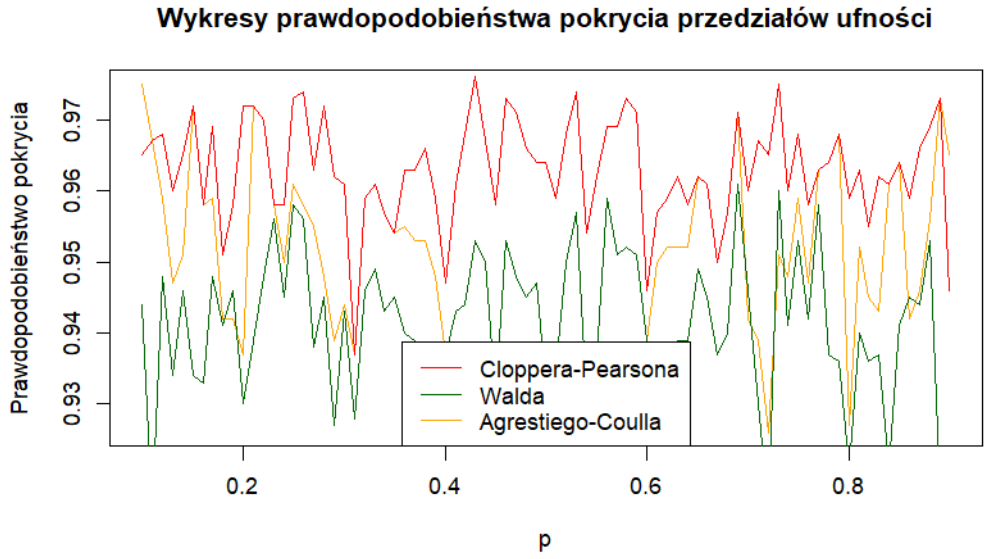
Rysunek 44: Wykresy prawdopodobieństwa pokrycia w zależności od p dla rozmiaru próby $n = 30$.

Dla próby $n = 30$ z powyższych wykresów (rysunek 43, 44) możemy odczytać, że najlepszym wyborem będzie przedział ufności Agrestiego-Coulla. Cechuje się on najkrótszą średnią długością przedziałów ufności, a jego prawdopodobieństwo pokrycia jest porównywalne z przedziałami Cloppera-Pearsona, z delikatną przewagą dla przedziałów Cloppera-Pearsona. Przedział Agrestiego-Coulla zachowuje się również lepiej przy wartościach krańcowych zmiennej p (w bliskiej okolicy $p = 0$ oraz $p = 1$) dla prawdopodobieństwa pokrycia, natomiast dzieje się to kosztem większej długości przedziałów od pozostałych typów.

Sprawdzimy jednak czy sytuacja się zmieni gdy zwiększymy długość próby na $n = 100$. Musieliśmy ograniczyć krańcowe przedziały do wartości $p \in (0.1, 0.9)$ aby wykresy były bardziej czytelne oraz by możliwe było wyciągniecie z nich odpowiednich wniosków.



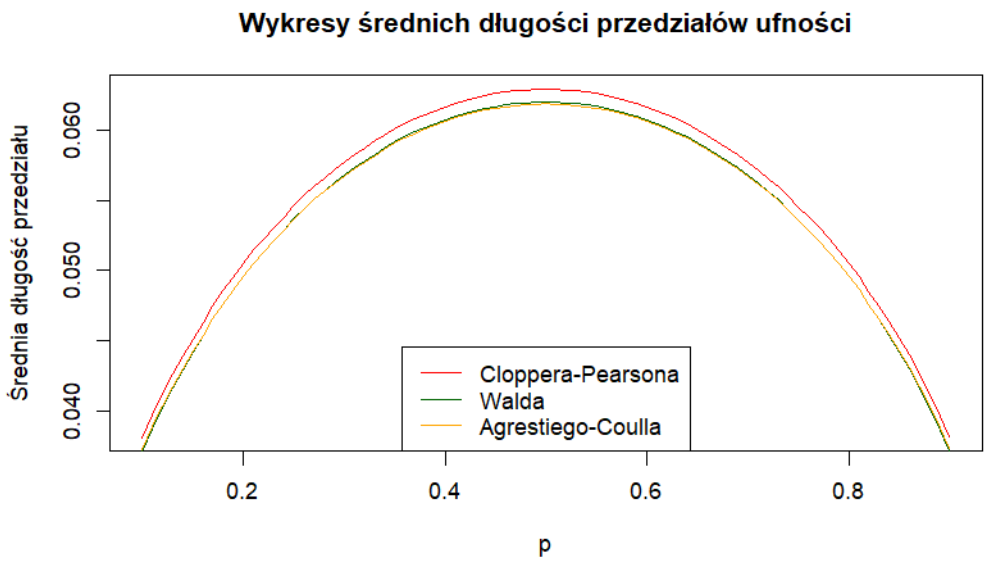
Rysunek 45: Wykresy średniej długości przedziałów ufności w zależności od p dla rozmiaru próby $n = 100$.



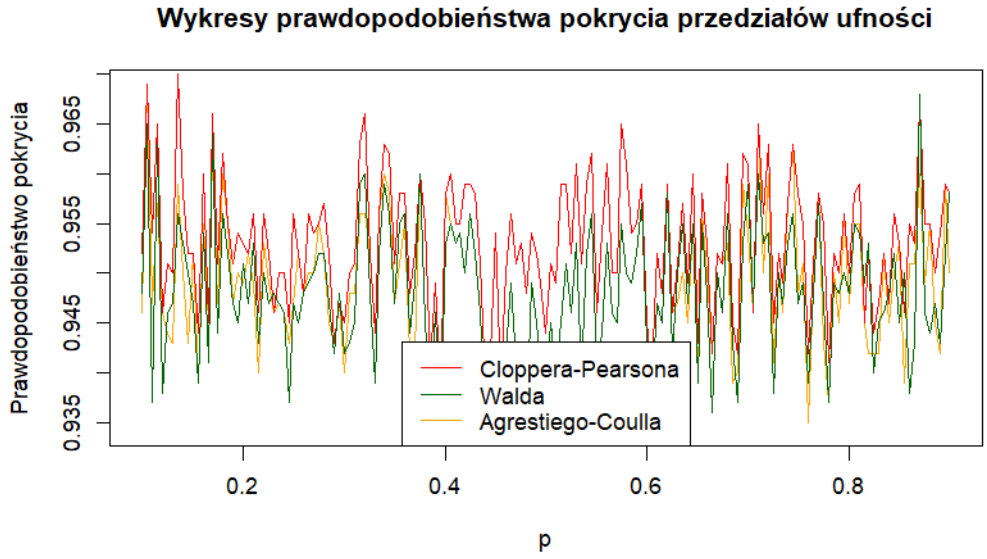
Rysunek 46: Wykresy prawdopodobieństwa pokrycia w zależności od p dla rozmiaru próby $n = 100$.

Przy zwiększonym rozmiarze próby mamy już nieco większy dylemat. Podobnie jak poprzednio przedział Cloppera-Pearsona posiada największe prawdopodobieństwo pokrycia oraz największą długość przedziału, z czego druga charakterystyka jest dla nas wysoce niepożądana. Najlepszym wyborem będzie ponownie przedział Agrestiego-Coulla, mimo że różnica między nim a przedziałami Walda jest już zdecydowanie mniejsza. Przedziały Walda w wartościach około $p = 0.2, p = 0.8$ zaczynają być najkrótsze, jednakże wiąże się to ze spadkiem prawdopodobieństwa pokrycia na rzecz innych typów przedziałów.

Ostatecznego porównania dokonamy dla próby o wielkości $n = 1000$.



Rysunek 47: Wykresy średniej długości przedziałów ufności w zależności od p dla rozmiaru próby $n = 1000$.



Rysunek 48: Wykresy prawdopodobieństwa pokrycia w zależności od p dla rozmiaru próby $n = 1000$.

Dochodzimy do konkluzji, że im większa długość próby, tym bardziej przedziały Walda oraz Agrestiego-Coulla się pokrywają. Dla próby $n = 1000$ ich średnia długość jest niemalże identyczna. Prawdopodobieństwo pokrycia wszystkich typów jest do siebie bardzo zbliżone. Ze względu na długość przedziału najlepiej jest natomiast wybrać przedział Walda lub Agrestiego-Coulla, ponieważ są one zdecydowanie krótsze niż przedział Cloppera-Pearsona.

6 Część czwarta

6.1 Zadania 10 i 11

Ostatnia część naszego raportu sprowadza się do testowania hipotez. W języku R służą do tego dwie funkcje. Dla pierwszych dwóch hipotez zastosujemy funkcję *binom.test*, a dla pozostałych - funkcję *prop.test*. Dla wszystkich testów przyjęliśmy poziom istotności $\alpha = 0.05$.

Funkcja *binom.test* przyjmuje 5 argumentów: x, n, p, alternative i conf.level, które odpowiednio oznaczają: liczbę sukcesów, liczność badanej populacji, prawdopodobieństwo podane w hipotezie, rodzaj hipotezy alternatywnej ("two.sided" - !=, "less" - <, "greater" - >) i poziom istotności (tutaj podajemy wartość $1 - \alpha$, czyli w naszym wypadku 0.95).

Funkcja *prop.test* ma podobne argumenty do tych z funkcji *binom.test* z tą różnicą, że x, n i p są wektorami, które zawierają liczby sukcesów w każdej z prób.

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

   Będziemy rozpatrywać następujące hipotezy:

   $$H_0 : p = 0.5$$

   $$H_1 : p \neq 0.5$$

   Dla tego testu kod wygląda jak poniżej:

   ```
   binom.test(71, 200, p = 0.5, alternative = "two.sided",
   conf.level = 0.95).
   ```

   Po jego realizacji otrzymujemy:

   - p-wartość: $4.973 * 10^{-5}$,
   - 95-procentowy przedział ufności: $(0.289, 0.426)$,
   - prawdopodobieństwo sukcesu: 0.355.

   P-wartość mniejsza od poziomu istotności α może świadczyć o odrzuceniu hipotezy zerowej na korzyść hipotezy alternatywnej, stąd prawdopodobieństwo tego, że w korporacji pracuje kobieta, nie jest równe 0.5.

2. Prawdopodobieństwo, że pracownik jest zadowolony ze swojego wynagrodzenia jest większe bądź równe 0.8.

   W tym przypadku rozpatrujemy poniższe hipotezy:

   $$H_0 : p \geqslant 0.8,$$

   $$H_1 : p < 0.8.$$

   Do weryfikacji hipotezy zerowej zastosujemy następujące kody odpowiednio dla zmiennych $W1$ i $W2$:

   ```
   binom.test(106, 200, p = 0.8, alternative = "less",
   conf.level = 0.95),

   binom.test(107, 200, p = 0.8, alternative = "less",
   conf.level = 0.95).
   ```

   Jak możemy zauważyć, wyniki dla obu testów będą niemal identyczne. W każdym z nich otrzymaliśmy p-wartość mniejszą od $2.2 * 10^{-16}$, co przekłada się na odrzucenie hipotezy H_0 i przyjęcie H_1, a zatem prawdopodobieństwo znalezienia pracownika zadowolonego ze swojego wynagrodzenia jest mniejsze niż 0.8.

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

   Analizowane hipotezy to:

   $$H_0 : p_k = p_m,$$
   $$H_1 : p_k \neq p_m,$$

   gdzie p_k oznacza prawdopodobieństwo znalezienia kobiety na stanowisku kierowniczym, a p_m to prawdopodobieństwo tego, że mężczyzna znajduje się na takim stanowisku.

   Program testujący hipotezę H_0 przyjmuje poniższą postać:

   ```
   prop.test(c(8,19), c(71,129), alternative = "two.sided",
   conf.level = 0.95)
   ```

   Otrzymaliśmy tutaj p-wartość równą w przybliżeniu 0.634, co pozwala nam przyjąć hipotezę zerową jako prawdziwą, zatem prawdopodobieństwo znalezienia kobiety na stanowisku kierowniczym jest równe prawdopodobieństwu znalezienia mężczyzny na takim stanowisku.

4. Prawdopodobieństwo, że kobieta jest zadowolona ze swojego wynagrodzenia jest równe prawdopodobieństwu, że mężczyzna jest zadowolony ze swojego wynagrodzenia.

   Testowanie przeprowadzimy dla obu badań opinii nt. wynagrodzenia. Hipotezy (w obu pytaniach) przyjmują postacie:

   $$H_0 : p_k = p_m,$$
   $$H_1 : p_k \neq p_m,$$

   gdzie p_k oznacza prawdopodobieństwo, że kobieta jest zadowolona ze swojego wynagrodzenia, a p_m to prawdopodobieństwo tego, że mężczyzna cieszy się ze swojej wypłaty.

   Dla pierwszego badania (zmienna W1) realizujemy poniższy kod:

   ```
   prop.test(c(36,70), c(71,129), alternative = "two.sided",
   conf.level = 0.95),
   ```

   a dla drugiego (zmienna W2) następujący:

   ```
   prop.test(c(37,70), c(71,129), alternative = "two.sided",
   conf.level = 0.95).
   ```

   Dla tych testów p-wartości wynoszą odpowiednio 0.738 i 0.886. Oznacza to, że w obu przypadkach możemy uznać odpowiednie hipotezy zerowe za prawdziwe. Stąd prawdopodobieństwo, że kobieta jest zadowolona ze swojego wynagrodzenia, było równe prawdopodobieństwu, że mężczyzna jest usatysfakcjonowany swoją wypłatą w obu ankietach.

5. Prawdopodobieństwo, że kobieta pracuje w dziale obsługi kadrowo-płacowej jest większe lub równe prawdopodobieństwu, że mężczyzna pracuje w dziale obsługi kadrowo-płacowej.

   Określamy hipotezy H_0 i H_1.

   $$H_0 : p_k \geqslant p_m,$$
   $$H_1 : p_k < p_m,$$

   p_k oznacza prawdopodobieństwo, że kobieta jest zatrudniona w dziale obsługi kadrowo-płacowej a p_m to prawdopodobieństwo tego, że mężczyzna pracuje w tymże dziale.

   Do zrealizowania testu użyjemy kodu:

   ```
   prop.test(c(23,3), c(71,129), alternative = "less",
   conf.level = 0.95).
   ```

   W jego wyniku otrzymujemy p-wartość równą 1, co oznacza, że hipoteza zerowa jest prawdziwa, zatem prawdopodobieństwo znalezienia kobiety w dziale obsługi kadrowo-płacowej jest większe niż prawdopodobieństwo znalezienia tam mężczyzny.