Analiza Danych Ankietowych
Sprawozdanie 1

Natalia Lach 262303, Alicja Myśliwiec 262275

Matematyka Stosowana
Wydział Matematyki Politechniki Wrocławskiej

2023

Spis treści

1. Wprowadzenie

Sprawozdanie dotyczy nierzeczywistych danych, wygenerowanych na rzecz niniejszej pracy. Badania ankietowe według założenia zostały przeprowadzaone na losowo (ze zwracaniem) wybranych dwustu pracownikach pewnej wielkiej korporacji. Poniżej zestawiono przedstawione im pytania.

1. (D) Pracuję w
   — (Z) w dziale zaopatrzenia,
   — (P) w dziale produkcyjnym,
   — (S) w dziale sprzedaży (w tym marketingu),
   — (O) w dziale obsługi kadrowo-płacowej.

2. (S) Pracuję na stanowisku kierowniczym
   — (1) tak,
   — (0) nie.

3. (A1 i A2) Atmosfera w miejscu pracy jest bardzo dobra
   — (-2) zdecydowanie się nie zgadzam,
   — (-1) nie zgadzam się,
   — (0) trudno powiedzieć,
   — (1) zgadzam się,
   — (2) zdecydowanie się zgadzam.

4. (W1 i W2) Jestem zadowolona/y ze swojego wynagrodzenia
   — (-2) zdecydowanie się nie zgadzam,
   — (-1) nie zgadzam się,
   — (1) zgadzam się,
   — (2) zdecydowanie się zgadzam.

5. (P) Płeć:
   — (K) kobieta,
   — (M) mężczyzna.

6. (Wiek) Wiek:
   — (1) do 25 lat,
   — (2) od 26 do 35 lat,
   — (3) od 36 do 50 lat,
   — (4) powyżej 50 lat.

7. (Wyk) Wykształcenie:
   — (1) zawodowe,
   — (2) średnie,
   — (3) wyższe.

Oznaczenia znajdujące się przy powyższych pytaniach będą odtąd używane we wszelkich tabelach w kolejnych częściach sprawozdania i będą odnosiły się do wymienionych odpowiedzi i zmiennych.

Zmienne A1 i A2 oraz W1 i W2, są odpowiednio ocenami atmosfery i wynagrodzenia w pierwszym i drugim badanym okresie. Niektóre z używanych funkcji biblioteki *likert* języka R, między innymi *summary* oraz wykres typu *heat* (zadanie w sekcji 3), transformują skalę likerta zastosowaną powyżej dla zmiennych A1 i A2. Otrzymane odpowiedzi odpowiadają w ten sposób nieco innym cyfrom, w podany poniżej sposób.

— (1) zdecydowanie się nie zgadzam,
— (2) nie zgadzam się,
— (3) trudno powiedzieć,
— (4) zgadzam się,
— (5) zdecydowanie się zgadzam.

Dzięki takiej transformacji, wynik średniej z funkcji *summary*, wynoszący 3, odpowiada równowadze ocen pozytywnych jak i negatywnych. W oryginalnej skali, w tym wypadku, średnia wynosiłaby 0. Owa informacja będzie użyta podczas analizy w kolejnych częściach sprawozdania (tj. w sekcji 3).

2. Część I

2.1. Zadanie 1.

Celem zadania było sporządzenie tablic liczności dla zmiennych $A1$ oraz $W1$, biorąc pod uwagę wszystkie dane, jak również w podgrupach ze względu na zmienne opisujące dział (D), płeć (P) oraz wykształcenie pracownika (Wyk). W celu swobodnego poruszania się po danych, zostały one sformatowane w następujący sposób (zgodny z oznaczeniami w sekcji 1).

```
## Wczytanie danych
data <- read.csv2("personel.csv", header=FALSE)
## Opisanie kolumn
names(data) <- c('D','S','A1','A2','W1','W2','P','Wiek','Wyk')
```

	D	S	A1	A2	W1	W2	P	Wiek	Wyk
	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>	<fct>
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

Rys. 1: head(data)

2.1.1. Dla A1 i W1

W celu wywołania tabeli dla zmiennej A1 skorzystamy z poniższego kodu

```
## Klasyczne wywolanie funkcji
mutate(count(data, A1), prop=n/sum(n))
## Wywolanie przy pomocy operatora pipe %<%
data %>% count(A1) %>% mutate(prop = n/sum(n))
```

A1	n	procent
-2	14	0.070
-1	17	0.085
0	40	0.200
1	100	0.500
2	29	0.145

Tab. 1: Tabela zmiennej A1

W1	n	procent
-2	74	0.370
-1	20	0.100
1	2	0.010
2	104	0.520

Tab. 2: Tabela zmiennej W1

2.1.2. Dla A1 ze względu na pozostałe zmienne

Aby sporządzić tablice w podgrupach, skorzystamy z funkcji *filter*.

```
data %>% filter(Wyk=="1") %>% count(A1) %>% mutate(prop=n/sum(n))
```

— Ze względu na zmienną *D*

A1	n	procent
-2	2	0.06451613
-1	2	0.06451613
0	5	0.16129032
1	19	0.61290323
2	3	0.09677419

Tab. 3: Tabela zmiennej A1 dla D = "Z"

A1	n	procent
-2	9	0.09183673
-1	10	0.10204082
0	17	0.17346939
1	51	0.52040816
2	11	0.11224490

Tab. 4: Tabela zmiennej A1 dla D = "P"

A1	n	procent
-2	3	0.06666667
-1	3	0.06666667
0	14	0.31111111
1	15	0.33333333
2	10	0.22222222

Tab. 5: Tabela zmiennej A1 dla D = "S"

A1	n	procent
-2	0	0
-1	2	0.07692308
0	4	0.15384615
1	15	0.57692308
2	5	0.19230769

Tab. 6: Tabela zmiennej A1 dla D = "O"

— Ze względu na zmienną P

A1	n	procent
-2	3	0.04225352
-1	7	0.09859155
0	14	0.19718310
1	36	0.50704225
2	11	0.15492958

Tab. 7: Tabela zmiennej A1 dla P = "K"

A1	n	procent
-2	11	0.08527132
-1	10	0.07751938
0	26	0.20155039
1	64	0.49612403
2	18	0.13953488

Tab. 8: Tabela zmiennej A1 dla P = "M"

— Ze względu na zmienną Wyk

A1	n	procent
-2	5	0.12195122
-1	6	0.14634146
0	8	0.19512195
1	19	0.46341463
2	3	0.07317073

Tab. 9: A1 dla Wyk = 1

A1	n	procent
-2	5	0.03571429
-1	10	0.07142857
0	26	0.18571429
1	75	0.53571429
2	24	0.17142857

Tab. 10: A1 dla Wyk = 2

A1	n	procent
-2	4	0.21052632
-1	1	0.05263158
0	6	0.31578947
1	6	0.31578947
2	2	0.10526316

Tab. 11: A1 dla Wyk = 3

2.1.3. Dla W1 ze względu na pozostałe zmienne

— Ze względu na zmienną D

W1	n	procent
-2	9	0.29032258
-1	3	0.09677419
1	0	0
2	19	0.61290323

Tab. 12: Tabela zmiennej W1 dla D = "Z"

W1	n	procent
-2	37	0.6755102
-1	11	0.11224490
1	1	0.01020408
2	49	0.50000000

Tab. 13: Tabela zmiennej W1 dla D = "P"

W1	n	procent
-2	20	0.44444444
-1	2	0.04444444
1	0	0
2	23	0.51111111

Tab. 14: Tabela zmiennej W1 dla D = "S"

W1	n	procent
-2	8	0.30769231
-1	4	0.15384615
1	1	0.03846154
2	13	0.50000000

Tab. 15: Tabela zmiennej W1 dla D = "O"

— Ze względu na zmienną P

W1	n	procent
-2	25	0.35211268
-1	10	0.14084507
1	1	0.01408451
2	35	0.49295775

Tab. 16: Tabela zmiennej W1 dla P = "K"

W1	n	procent
-2	49	0.379844961
-1	10	0.077519380
1	1	0.007751938
2	69	0.534883721

Tab. 17: Tabela zmiennej W1 dla P = "M"

— Ze względu na zmienną Wyk

W1	n	procent
-2	20	0.48780488
-1	3	0.07317073
1	0	0
2	18	0.43902439

Tab. 18: W1 dla Wyk = 1

W1	n	procent
-2	45	0.3214286
-1	17	0.1214286
1	0	0
2	78	0.5571429

Tab. 19: W1 dla Wyk = 2

W1	n	procent
-2	9	0.4736842
-1	0	0
1	2	0.1052632
2	8	0.4210526

Tab. 20: W1 dla Wyk = 3

2.2. Zadanie 2.

Tabela wielodzielcza przedstawia rozkład łączny dwóch wybranych zmiennych. Przykładowo, tabele uwzględniającą zmienną W1 i P Można uzyskać następująco.

```
## sposob 1
ftable(data, col.vars="W1", row.vars="P")
## sposob 2
structable(W1~P, data) %>% addmargins()
```

	-2	-1	1	2	Suma
K	25	10	1	35	71
M	49	10	1	69	129
Suma	74	20	2	104	200

Tab. 21: W1 i P

	-2	-1	1	2	Suma
0	64	18	0	91	173
1	10	3	2	13	27
Suma	74	20	2	104	200

Tab. 22: W1 i S

	-2	-1	0	1	2	Suma
O	0	2	4	15	5	26
P	9	10	17	51	11	98
S	3	3	14	15	10	45
Z	2	2	5	19	3	31
Suma	14	17	40	100	29	200

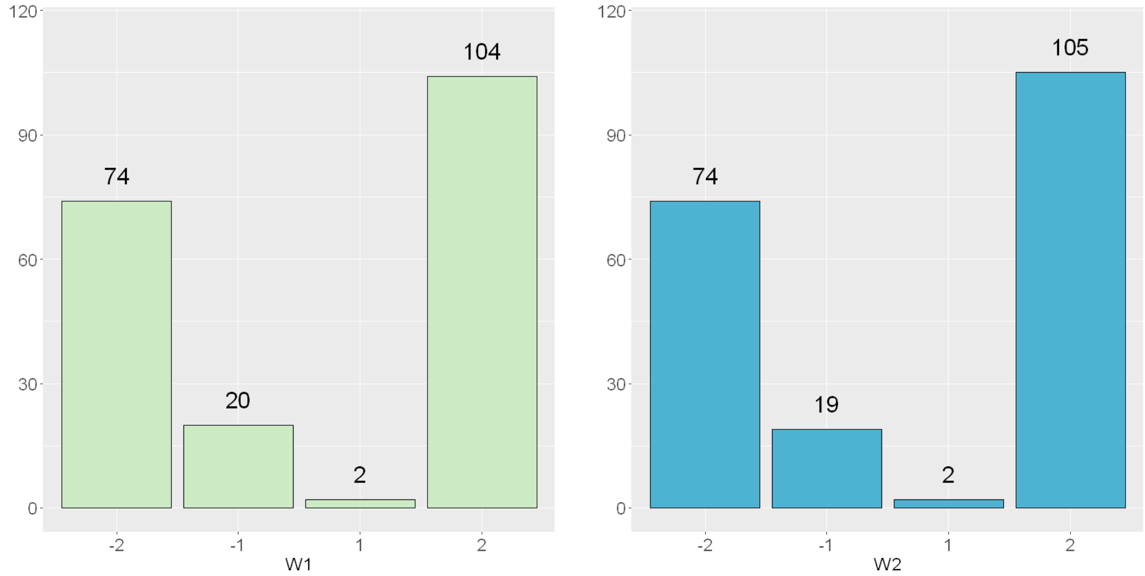
Tab. 23: A1 i D

2.3. Zadanie 3.

Przykładowe kody generowania wykresów dla zmiennej W1.

```
dane <- data %>% count(W1)
## Kod generujacy wykres slupkowy
ggplot(daneW1, aes(x=W1, y=n)) +
    geom_bar(stat="identity", color = "black", fill = '#ccebc5') +
    labs(x="W1", y = "") + coord_cartesian(ylim = c(0, 120)) +
    geom_text(aes(label=n), vjust=-1, color="black", size=7) +
    theme(axis.text=element_text(size=15), axis.title=element_text(size=15))
## Kod generujacy wykres kolowy
ggplot(daneW1, aes(x="", y=n, fill=W1)) + geom_bar(width=1, stat="identity") +
    coord_polar(theta="y", start=45) + geom_text(aes(label = n),
    position = position_stack(vjust=0.6), size = 5) +
    labs(x="", y="", fill="Odpowiedzi") + scale_fill_brewer(palette="GnBu") +
    theme(axis.line = element_blank(), axis.text = element_blank()) +
    theme(legend.key.size = unit(2, 'cm'),
    legend.title = element_text(size=15), legend.text = element_text(size=15))
```

2.3.1. Wykresy słupkowe



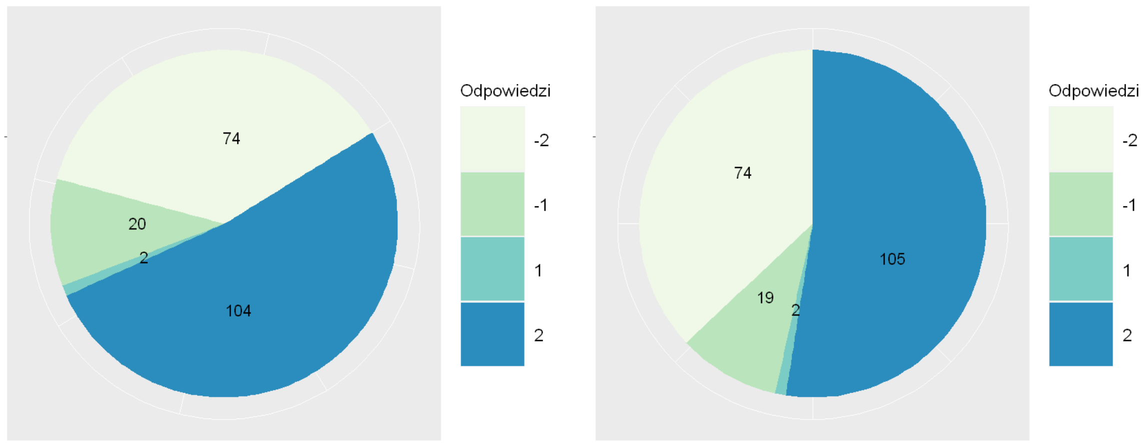
(a) Wykres zmiennej W1.

(b) Wykres zmiennej W2.

Rys. 2: Wykresy słupkowe.

Wykresy na rysunku 2 przedstawiają odpowiednio ilości poszczególnych odpowiedzi na pytanie dotyczące zadowolenia z wynagrodzenia w dwóch badanych okresach. Odpowiedzi są zaskakująco identyczne. Może wynikać to z faktu, iż dane są nierzeczywiste. Odpowiedzi *Zdecydowanie się zgadzam* (2) jest sumarycznie więcej niż pozostałych.

2.3.2. Wykresy kołowe



(a) Wykres zmiennej W1.

(b) Wykres zmiennej W2.

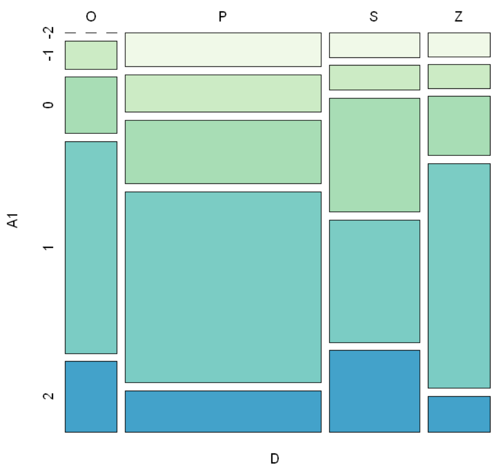
Rys. 3: Wykresy kołowe.

Na wykresach kołowych na rysunku 3 zdecydowanie wyraźniej zauważalna jest przewaga odpowiedzi (2). Kolejną co do częstości jest natomiast odpowiedź *Zdecydowanie się nie zgadzam* (-2). Dwie pozostałe mają znikomy udział, co świadczy o zdecydowaniu w odpowiedziach badanych pracowników.

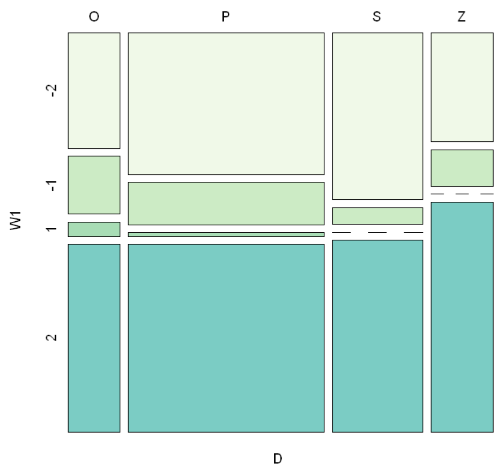
2.4. Zadanie 4.

Generowanie wykresu mozaikowego dla pary zmiennych D oraz A1.

```
mosaicplot(~D+A1,data,main="",color=(brewer.pal(n=6,name="GnBu")),cex.axis=1)
```



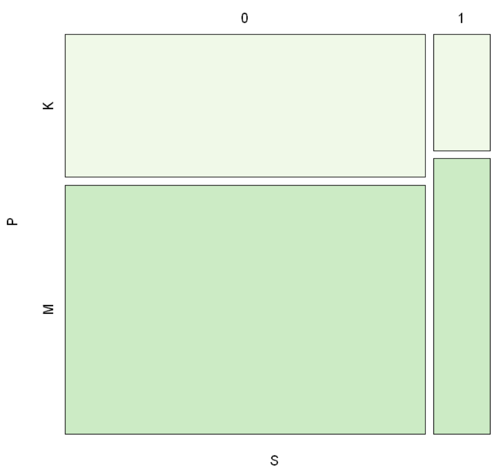
Rys. 4: Para zmiennych D i A1



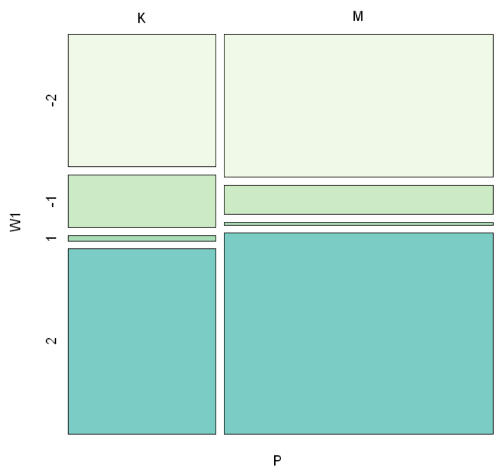
Rys. 5: Para zmiennych D i W1

Z rysunku 4 można odczytać, iż w zależności od pracy w konkretnym dziale zmienia się udział procentowy odpowiedzi oceniających dobrą atmosferę pracy. Dział obsługi kadrowo-płacowej ma najwięcej pozytywnych ocen oraz żadnej odpowiedzi *"Zdecydowanie się nie zgadzam"*. Najwięcej negatywnych opinii procentowo można zauważyć w dziale produkcji, za to dział sprzedaży ma najwięcej odpowiedzi neutralnych. W ogólnym rozrachunku atmosfera pracy w każdym z działów jest na wysokim poziomie.

Analizując jednak zależność działu względem zadowolenia z zarobków, przedstawioną na rysunku 5, zauważalne jest już dużo większy udział odpowiedzi negatywnych. Warto zaznaczyć, że zdecydowana część pracowników niezależnie od działu jest bardzo zadowolona z zarobków, w szczególności zaopatrzeniowcy. Procentowy udział odpowiedzi pośrednich jest znikomy.



Rys. 6: Para zmiennych S i P



Rys. 7: Para zmiennych P i W1

Rysunek 6 przedstawia obsadzenie stanowisk kierowniczych w zależności od płci. Można zauważyć, że zdecydowaną część tych stanowisk zajmują mężczyźni (około dwukrotnie więcej niż kobiety). Warto jednak zwróćić uwagę, iż większość zatrudnionych osób w korporacji, to właśnie mężczyźni.

Wspomniana dominacja płci męskiej w omawianym zakładzie pracy, nie wpłyneła jednak na duże zróżnicowanie zadowolenia z zarobków w przypadku obu płci, co pokazane jest na rysunku 7. Zauważalne jest mniejsze zdecydowanie kobiet w udzielanych odpowiedziach, jednak w obu przypadkach góruje odpowiedź *"Zdecydowanie się zgadzam"*.

3. Część II

3.1. Zadanie 5.

W celu wykonania zadania posłużymy się bazą danych *mtcars* dostępną w języku R.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1

Rys. 8: mtcars

```
## bez zwracania
mtcars[sample(1:nrow(mtcars), nrow(mtcars)/10),]
## ze zwracaniem
mtcars[sample(1:nrow(mtcars), nrow(mtcars)/10), replace=TRUE]
```

Powyższy kod zwrócił wyniki w postaci tabeli przedstawionych poniżej.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2

(a) Losowanie bez zwracania

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Ford Pantera L	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8

(b) Losowanie ze zwracaniem

Rys. 9: Losowanie mtcars

3.2. Zadanie 6.

Fragment kodu realizujący podane zadanie przedstawiono poniżej.

```
##obsluga bledu w danych
data$A2[data$A2 == 11] = 1

##zfaktoryzowanie wartosci
data <- data %>% mutate(across(D:Wyk, as.factor))

##stworzenie obiektow klasy likert - cala probka
likt <- likert(data.frame("Atmosfera_1" = data$A1,
    "Atmosfera_2" = data$A2))
##stworzenie obiektow klasy likert - z podzialem na podgrupy
likt2 <- likert(data.frame("Atmosfera_1" = data$A1,
    "Atmosfera_2" = data$A2), grouping=data$D)

##summary
summary(likt)
##wykresy
plot(likt)
plot(likt, type="density")
plot(likt, type="heat")
```

W niniejszym zadaniu zastosowane zostały inne oznaczenia dla zmiennych A1 i A2, odpowiednio $Atmosfera1$ i $Atmosfera2$.

3.2.1. Cała próba

Tabela przedstawiająca *summary* całej próby dotyczącej oceny atmosfery w miejscu pracy w przetransformowanej przez język R skali Likerta[1].

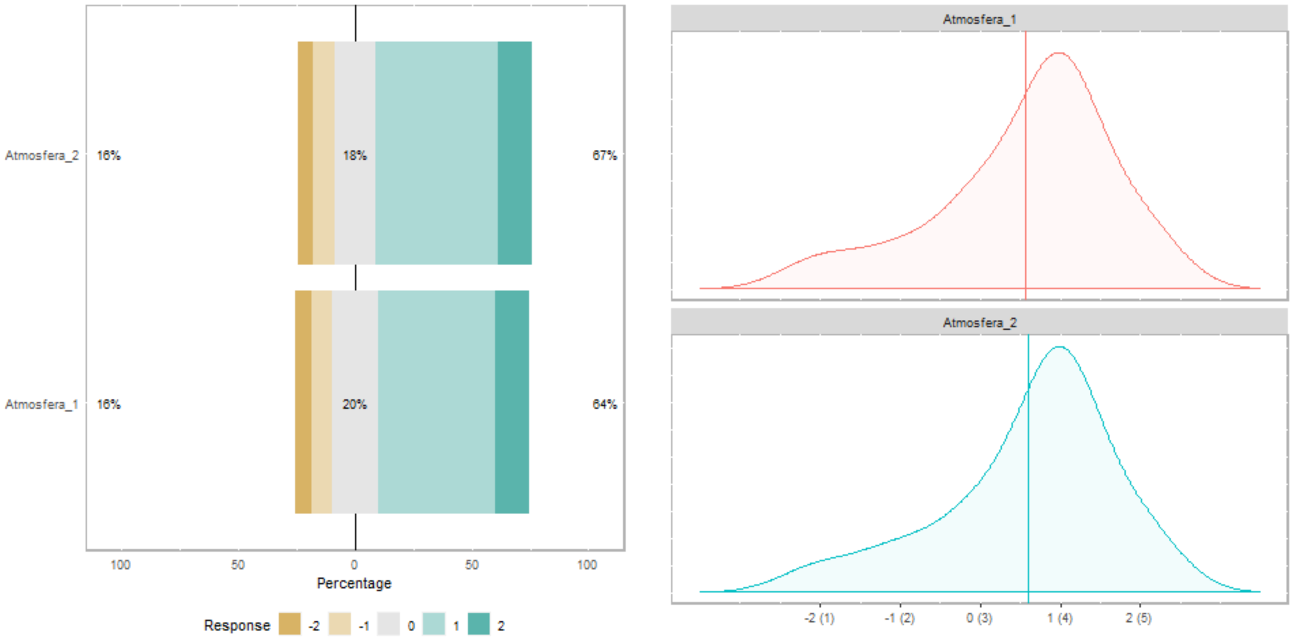
	Item	low	neutral	high	mean	sd
2	Atmosfera_2	15.5	17.5	67.0	3.600	1.041838
1	Atmosfera_1	15.5	20.0	64.5	3.565	1.063688

Rys. 10: Podsumowanie całej próby.

Z tabeli na rysunku 10 wynika, że atmosfera w miejscu pracy była lepsza w drugim badaniu. Wskazuje na to nieco wyższa średnia oraz mniejsze odchylenie standardowe dla zmiennej $Atmosfera1$. Najmniejszy udział procentowy w obu przypadkach mają oceny negatywne (tylko 15.5%)

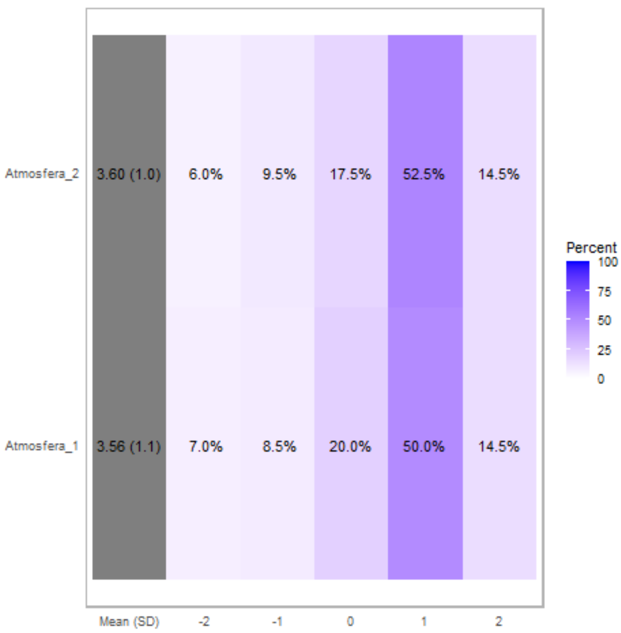
[1] Skala ta została przedstawiona na stronie 3

Poniżej zwizualizowane zostały rozkłady danych przy pomocy wykresów, między innymi gęstości.



(a) Rozkład procentowy. (b) Gęstość.

Rys. 11: Gęstość oraz rozkład procentowy całej próby.



Rys. 12: Wykres typu heat całej próby.

Wykresy z rysunków 11 i 12 dobrze ilustrują podobieństwo rozkładów danych dotyczących atmosfery w miejscu pracy dla dwóch badanych okresów. Różnice są niemal niezauważalne na wykresie gęstości i rozkładzie procentowym. Wykres *heat* ujawnia zmianę ilości ocen składających się na ogólny wynik. Atmosfera badania w drugim okresie zyskała mniej silnie negatywnych i neutralnych ocen. Największy udział w ankiecie mają odpowiedzi pozytywne. W przypadku $Atmosfery1$, ich udział jest o 2.5% niższy niż w przypadku $Atmosfery2$. Niemniej jednak, to $Atmosfera2$ zyskała wyższą średnią ogólną, co zostało zauważone wcześniej.

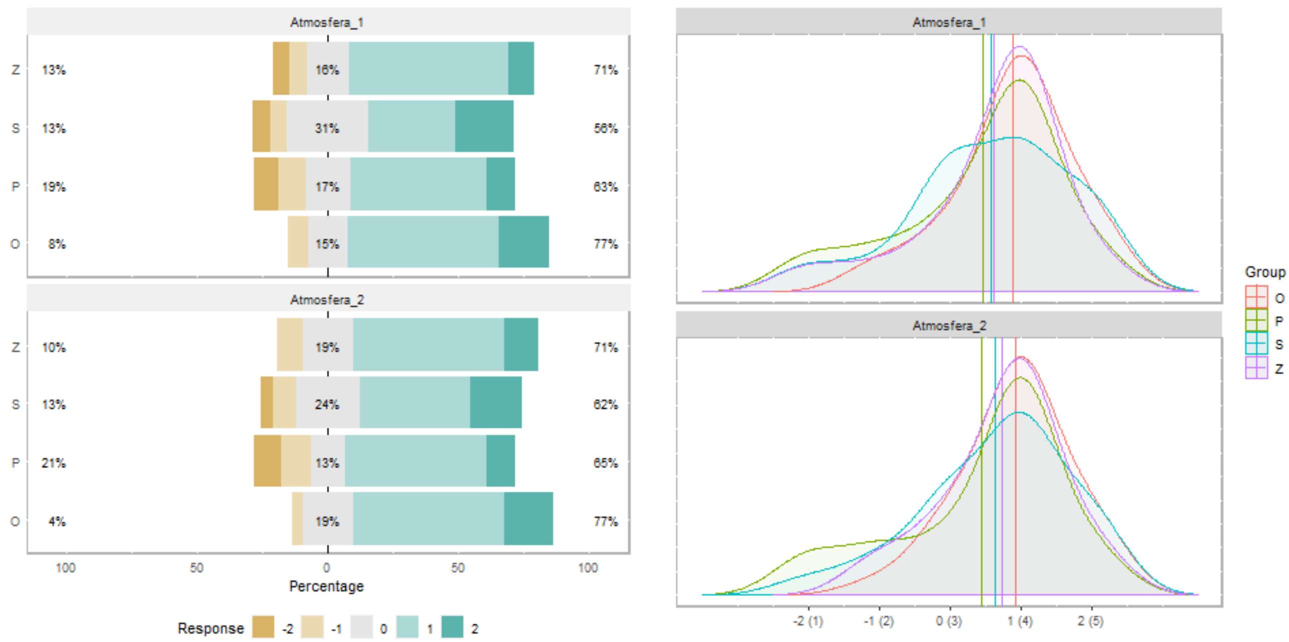
3.2.2. Podział próby względem działu.

Teraz próba została pogrupowana ze względu na działy w rozważanej firmie. Poniżej podsumowanie danych przy pomocy *summary*.

	Group	Item	low	neutral	high	mean	sd
1	O	Atmosfera_1	7.692308	15.38462	76.92308	3.884615	0.8161825
2	O	Atmosfera_2	3.846154	19.23077	76.92308	3.923077	0.7442084
3	P	Atmosfera_1	19.387755	17.34694	63.26531	3.459184	1.1138155
4	P	Atmosfera_2	21.428571	13.26531	65.30612	3.448980	1.1498314
5	S	Atmosfera_1	13.333333	31.11111	55.55556	3.577778	1.1178081
6	S	Atmosfera_2	13.333333	24.44444	62.22222	3.644444	1.0478453
7	Z	Atmosfera_1	12.903226	16.12903	70.96774	3.612903	0.9891889
8	Z	Atmosfera_2	9.677419	19.35484	70.96774	3.741935	0.8151786

Rys. 13: Podsumowanie próby względem działu.

Tym razem wyniki z tabeli na rysunku 13 są mocniej zróżnicowane. Każdy z działów ma inną średnią. Wyniki wskazują na to, że najbardziej zadowoleni są pracownicy z działu obsługi kadrowo-płacowej, a najmniej z działu produkcyjnego. Porównując wyniki w obrębie grup, nie widać jednoznacznie, że to atmosfera w drugim okresie badania jest wyższa. Zaprzecza temu wyższy osiągnięty wynik dla pracowników z działu produkcyjnego w pierwszym okresie badania. Grupy mają także dość mocno różniące się od siebie rozkłady procentowe ocen oraz odchylenia standardowe. Ponownie grupami odstającymi od reszty są działy obsługi kadrowo-płacowej i produkcyjny z zanotowanymi odpowiednio najniższym i najwyższym odchyleniem standardowym.



(a) Rozkład procentowy (b) Gęstość

Rys. 14: Gęstość oraz rozkład procentowy próby względem działu.

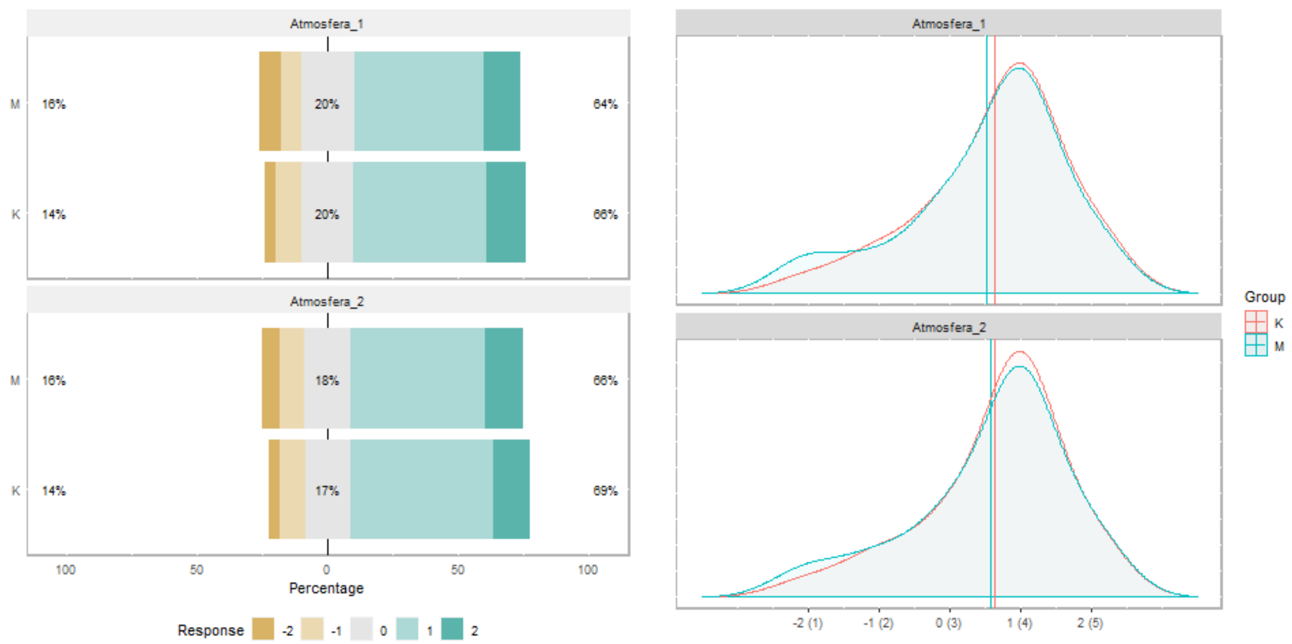
Analizując wykresy na rysunku 14 można zauważyć większe zmiany w rozkładach ocen niż w przypadku całej próby. Wizualnie najbardziej zmienił się wykres gęstości dla działu sprzedaży. Dla $Atmosfery1$ ma on znacznie różniący się od reszty grup kształt. Jest bardziej spłaszczony, co oznacza, że w w pierwszym okresie badania wystąpiła porównywalna ilośc ocen neutralnych i pozytywnych. W pozostałej części przypadków to opinii *"Zgadzam się"* było najwięcej. Rozważana zmiana nastawienia jest również widoczna w rozkładzie procentowym danych. Różnice można także zauważyć w opiniach pracowników z działu zaopatrzenia. Nastąpiła tam poprawa atmosfery, biorąc pod uwagę brak ocen silnie niegatywnych w drugim okresie badania. Analizując wyniki w obrębie grup, podobnie jak wcześniej, to odpowiedzi pozytywne mają największy udział w całości ankiet.

3.2.3. Podział próby względem płci.

	Group	Item	low	neutral	high	mean	sd
1	K	Atmosfera_1	14.08451	19.71831	66.19718	3.633803	1.0034147
2	K	Atmosfera_2	14.08451	16.90141	69.01408	3.647887	0.9870387
3	M	Atmosfera_1	16.27907	20.15504	63.56589	3.527132	1.0974225
4	M	Atmosfera_2	16.27907	17.82946	65.89147	3.573643	1.0736561

Rys. 15: Podsumowanie próby z podziałem na płeć.

Wyniki przedstawione w powyższej tabeli nie są aż tak zróżnicowane jak w przypadku tabeli z rysunku 13. Średnie są podobne w obu grupach, jednak to u kobiet zauważyć można nieco lepszą opinię o atmosferze w pracy w obu badanych okresach. W obu grupach średni wynik nieznacznie się poprawiła w drugim okresie. Zauważyć można spadek ilości ocen neutralnych na rzecz ocen pozytywnych.



(a) Rozkład procentowy (b) Gęstość

Rys. 16: Gęstość oraz rozkład procentowy próby względem płci.

Analizując wykresy z powyższego rysunku 16, nie zauważamy większych różnic pomiędzy dwoma badanymi okresami, ani pomiędzy grupami. Dla $Atmosfery1$ ilość ocen neutralnych jest według rozkładu procentowego identyczna dla obu grup, rozróżnia je jedynie ilość opinii negatywnych względem pozytywnych - kobiety czują się lepiej w miejscu pracy, co też można było wnioskować po wyższej osiągniętej średniej. Dla $Atmosfery2$ sytuacja jest analogiczna, lecz widać spadek ocen neutralnych na rzecz pozytywnych, co także zostało zauważone wcześniej. Wszystkie te zmiany nie są jednak duże, wykres gęstości dla obu grup i zmiennych niewiele się zmienił.

3.3. Zadanie 7.

Zadanie zostało wykonane za pomocą następującej funkcji.

```
clopper_pearson <- function(x, n, conf=NULL) {
    ## sprawdzenie, czy podany zostal wektor danych,
    # czy suma oraz dlugosc proby
    if (is.null(conf)) {
        alfa <- 1 - n
        n <- length(x)
        x <- sum(x)
        cat(alfa, n, x)
    }
    else {
        alfa <- 1 - conf
        cat(alfa, n, x)
    }
    ## wyliczenie dolnych i gornych wartosci przedzialu
    # w zaleznosci od przypadku
    if (x == 0) {
        pd <- 0
        pg <- qbeta(1 - alfa/2, x + 1, n - x)
        return(c(pd, pg))
        cat(pd, pg)
    }
    else if (x == n) {
        pg <- n
        pd <- qbeta(alfa/2, x, n - x + 1)
        return(c(pd, pg))
        cat(pd, pg)
    }
    else {
        pd <- qbeta(alfa/2, x, n - x + 1)
        pg <- qbeta(1-alfa/2, x + 1, n - x)
        return(c(pd, pg))
        cat(pd, pg)
    }
}
```

W ramach przetestowania działania funkcji, wywołano poniższy kod.

```
x <- rbinom(1, 20, 0.4)
clopper_pearson(x, 20, 0.95)
binom.confint(x, 20, methods="exact")
```

	method	x	n	mean	lower	upper
1	exact	12	20	0.6	0.3605426	0.8088099

Rys. 17: Wynik przy użyciu wbudowanej funkcji

$\underline{p}$	$\overline{p}$
0.36054258	0.80880994

Tab. 24: Przedziały ufności C-P dla losowej próby rbinom.

Wynik funkcji języka R przedstawiony jest na rysunku 17, natomiast wynik autorskiej funkcji *clopper pearson* zawarto w tabeli na rysunku 24. Oznaczenia do tabeli są następujące:

— $\underline{p}$ - dolna wartość szukanego przedziału ufności - odpowiednik *lower*,
— $\overline{p}$ - górna wartość szukanego przedziału ufności - odpowiednik *upper*.

Powyższe oznaczenia będą odtąd używane także w kolejnych zadaniach i tabelach.

Jak można zauważyć, otrzymane wartości $\underline{p}$ i *lower*, a także $\overline{p}$ i *upper*, są niemal identyczne. Dzięki takiej informacji, możemy kontynuować pracę na danych, korzystając z funkcji *clopper pearson*.

3.3.1. Cała próba.

Poniższa tabela zawiera parametry, które wynikają z danych dotyczących zadowolenia z wynagrodzenia w pierwszym badanym okresie. Liczba sukcesów odpowiada ilości odpowiedzi *"Zgadzam się"* oraz *"Zdecydowanie się zgadzam"*. Długość próby to ilość wszystkich ankietowanych. Takie też oznaczenia będą dotyczyć wszelkich tabel w rozważanym zadaniu, także podczas ewentualnego podziału na grupy.

Parametry		Przedział ufności	
Długość próby	Liczba sukcesów	$\underline{p}$	$\overline{p}$
200	106	0.4583	0.6008

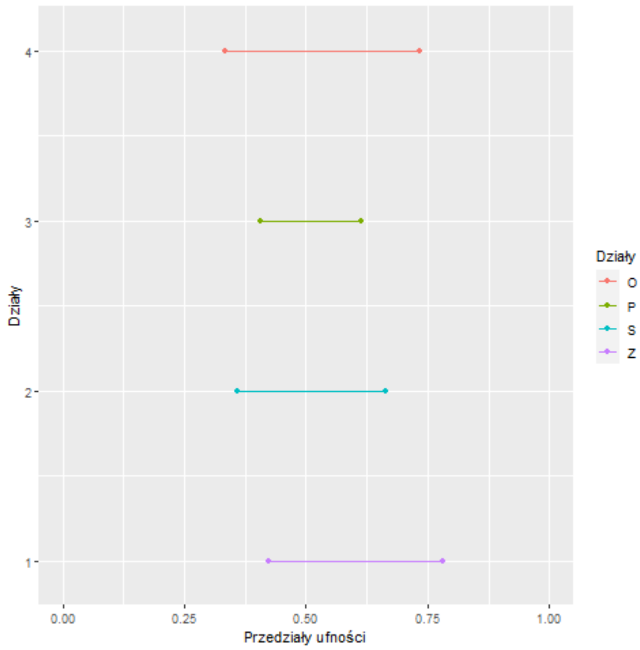
Tab. 25: Przedziały ufności C-P dla całej populacji.

Tabela 25 przedstawia otrzymane wyniki dotyczące dolnej i górnej wartości przedziału ufności Cloppera-Pearsona dla całej badanej populacji. Jest on dość wąski, to jest $\underline{p}$ i $\overline{p}$ są bliskie sobie.

3.3.2. Próba ze względu na dział.

Grupa	Parametry		Przedział ufności	
Dział	Długość próby	Liczba sukcesów	$\underline{p}$	$\overline{p}$
O	26	14	0.3337	0.7341
P	98	50	0.4072	0.6126
S	45	23	0.3577	0.6630
Z	31	19	0.4219	0.7815

Tab. 26: Przedziały ufności dla próby ze względu na dział



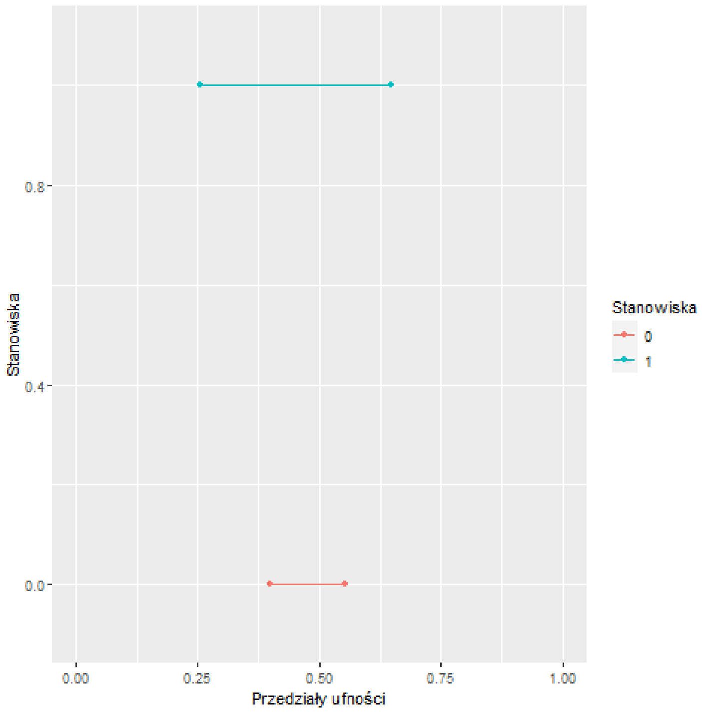
Rys. 18: Zwizualizowane przedziały ufności dla populacji ze względu na dział.

W tabeli 26 podane zostały odpowiednie parametry potrzebne do wywołania funkcji, obliczającej wartości przedziałów ufności. Dział P, to znaczy dział produkcyjny, stanowi największą grupę w firmie. Zanotował także najwięcej sukcesów, jednak zarazem uzyskał najmniejszy procent sukcesów wobec długości próby (około 51%). Jednocześnie otrzymana dla niego długość przedziału ufności jest najmniejsza spośród wszystkich działów, co zauważyć można na wykresie na rysunku 18. W ramach porównania, najdłuższym przedziałem ufności jest ten uzyskany dla działu O, czyli obsługi kadrowo-płacowej. Podobną długość zanotował dział zaopatrzenia, który jednocześnie uzyskał największy procent sukcesów (aż około 61%). Oba te działy miały jednak najmniejszą długość próby, co wpłynęło na większą długość przedziału ufności.

3.3.3. Próba ze względu na zajmowane stanowisko.

Grupa	Parametry		Przedział ufności	
Stanowisko	Długość próby	Liczba sukcesów	$\underline{p}$	$\overline{p}$
niekierownicze	173	91	0.4488	0.6023
kierownicze	27	15	0.3533	0.7452

Tab. 27: Przedziały ufności dla populacji ze względu na stanowisko.



Rys. 19: Zwizualizowane przedziały ufności dla populacji ze względu na stanowisko.

W przypadku podziału populacji ze względu na zajmowane stanowisko, zauważyć można dużą dysproporcję ilości stanowisk kierowniczych wobec zwykłych pracowników. Różnice w parametrach podanych w tabeli 27, znacznie wpływają na otrzymane długości przedziałów. Na rysunku 19 zauważyć można, że przedział ufności dla stanowisk niekierowniczych jest bardzo wąski. Uzyskane $\underline{p}$ i $\overline{p}$ są zbliżone wartością do odpowiadających im wyników dla całej populacji, otrzymanych w tabeli 25. Przyczyną są wprowadzone parametry - 173 jako długość próby jest bliska 200, a 91 sukcesów bliskie jest 106.

4. Część III

4.1. Zadanie 8.

4.1.1. Algorytm generowania oraz dowód

Wzór na funkcje charakterystyczną dla rozkładów dyskretnych wyraża się wzorem$\phi = \mathbb{E}e^{itX}$.

Dla zmiennej X z rozkładu dwumianowego tj. $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, mamy

$$\mathbb{E}e^{itX} = (1 - p + pe^{it})^n. \tag{1}$$

Niech

$$Y = \sum_{i=1}^{n} Z_i, \quad \text{gdzie} \begin{cases} P(Z=0) = 1-p \\ P(Z=1) = p \end{cases}$$

.

Wyprowadzenie dla zmiennej Z_i.

$$\mathbb{E}e^{itZ_i} = P(Z=0)e^{it\cdot 0} + P(Z=1)e^{it\cdot 1} = (1-p)e^{it\cdot 0} + pe^{it\cdot 1} = 1 - p + pe^{it} \tag{2}$$

Następnie dla Y.

$$\begin{aligned}\mathbb{E}e^{itY} &= \mathbb{E}e^{it\sum_{i=1}^{n} Z_i} = \mathbb{E}e^{itZ_1} \cdot e^{itZ_2} \cdot \ldots \cdot e^{itZ_n} = \mathbb{E}e^{itZ_1} \cdot \mathbb{E}e^{itZ_2} \cdot \ldots \cdot \mathbb{E}^{itZ_n} \\ &= \prod_{k=0}^{n} \mathbb{E}e^{itZ_i} = \prod_{k=0}^{n} 1 - p + pe^{it} = (1 - p + pe^{it})^n\end{aligned} \tag{3}$$

Teoretyczna funkcja charakterystyczna dla rozkładu dwumianowego jest równa wyznaczonej funkcji charakterstycznej zmiennej $Y = \sum_{i=1}^{n} Z_i$. Na tej podstawie można wygenerować zmienną z rozkładu dwumianowego, według poniższego algorytmu.

1. Wygenerowanie ciągu zmiennych losowych X_i, $i = 1{,}2, \ldots ,n$. Pojedynczą wartość X_i uzyskujemy, generując losową wartość v z przedziału $[0{,}1]$ i sprawdzając, czy $v < p$, gdzie p - ustalony wcześniej parametr. Jeśli warunek jest spełniony - zwracamy wartość 1, w przeciwnym wypadku - 0.
2. Zsumowanie otrzymanego ciągu, to jest wyliczenie $Y = \sum_{i=1}^{n} X_i$.
3. Zwrócenie wyliczonej wartości Y.

Aby stworzyć próbę losową złożoną z N wartości z rozkładu dwumianowego, powtarzamy powyższy algorytm N razy, tworząc wektor losowy. Kod jest przedstawiony poniżej.

4.1.2. Kod

```
 ## funkcja podstawowa, generujaca jedna probe Bernoulliego
gen_one <- function(p = 0.6) {
    if (runif(1) < p) {
        1
    }
    else {
        0
    }
}

##funkcja generujaca N prob z rozkladu dwumianowego o parametrach n i p
gen_binom <- function(N = 1000, n = 20, p = 0.6) {
    replicate(N, sum(replicate(n, gen_one(p))))
}

##wygenerowanie proby
proba <- gen_binom()
```

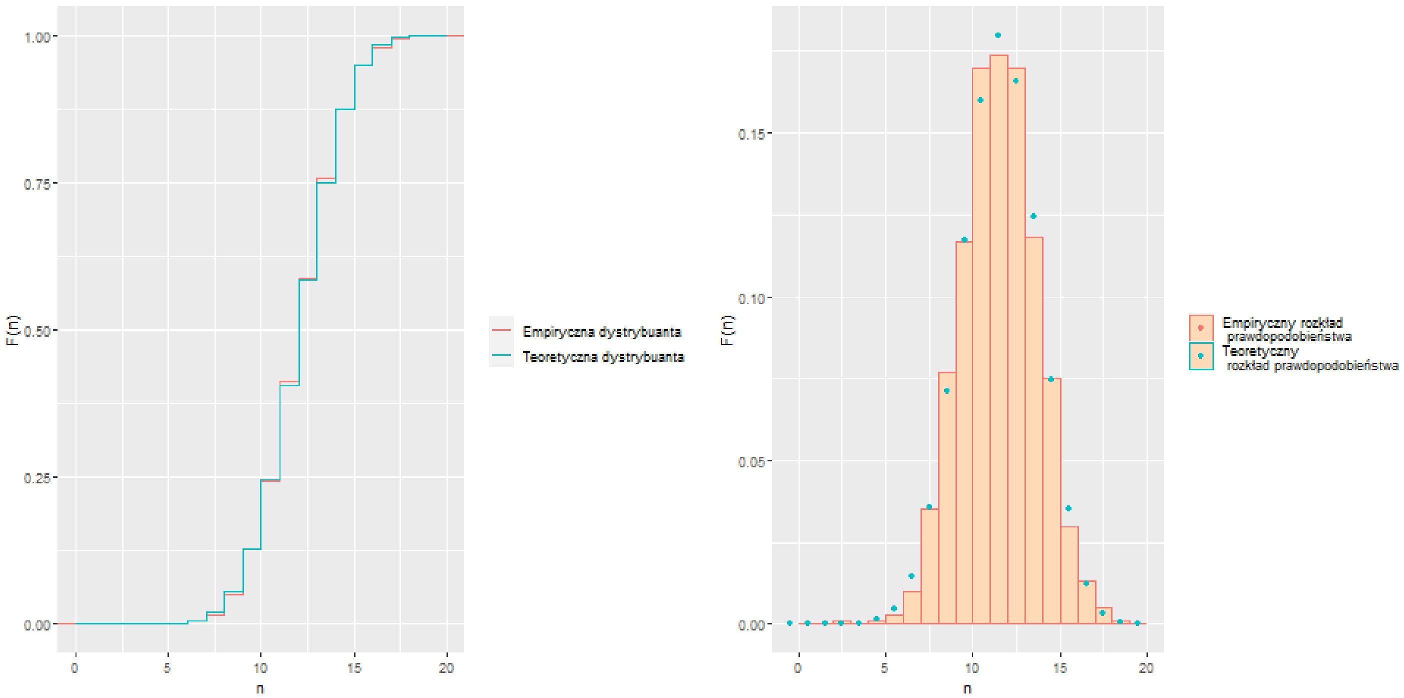
Wywołanie *mean(proba)* oraz *var(proba)* dało następujące wyniki.

Średnia	Wariancja
11.994	4.735

Tab. 28: Średnia i wariancja wygenerowanej próby.

Są one zgodne z teoretycznymi wartościami, czyli

$$\begin{cases}\mathbb{E}X = np = 12, \\ Var(X) = np(1-p) = 4.8.\end{cases} \tag{4}$$



(a) Dystrybuanta. (b) Rozkład prawdopodobieństwa.

Rys. 20: Dystrybuanta i rozkład prawdopodobieństwa dla wygenerowanego rozkładu dwumianowego.

Zwizualizowano także otrzymane dane za pomocą ich dystrybuanty i histogramy, jednocześnie porównując je z teoretycznymi wartościami tychże wykresów. Na rysunku 20 zauważyć można, że empiryczna i teoretyczna dystrybuanta są sobie bliskie i niemal się pokrywają. Wysokości słupków histogramu także oscylują wokół ich przewidywanych wartości. Nie są one już tak bliskie sobie jak w przypadku dystrybuanty. Prawdopododobnie po wygenerowaniu większej ilości próbek N, rozkłady pokryłyby się ze sobą. Ogólnie stwierdzić jednak można, że funkcja *gen_binom* faktycznie generuje realizacje z rozkładu dwumianowego.

4.2. Zadanie 9.

Funkcje do symulowania procentu pokrycia przedziałów ufności przedstawione są poniżej.

```
## funkcja sprawdzajaca czy dane p wpada do ustalonego przedzialu ufnosci
check_one <- function(n=20, p=0.6) {
    x <- binom.confint(rbinom(1, n, p), n, 0.95, "exact" )
    y <- binom.confint(rbinom(1, n, p), n, 0.95, "asymptotic" )
    z <- binom.confint(rbinom(1, n, p), n, 0.95, "bayes" )
    c(p<x$upper & p>x$lower, p<y$upper & p>y$lower, p<z$upper & p>z$lower)
}

##funkcja symulujaca N powtorzen Monte Carlo funkcji check_one
sym1 <- function(p, N = 30, n = 20) {
    rowMeans(replicate(N, check_one(n, p)))
}
```

```
## wygenerowanie wysymulowanych danych dla N=30, 100, 1000 dla trzech metod
p30 <- round(mapply(sym1, seq(0,1,0.1), N = 30), 3)

##dodatkowe zaokraglenie wynikow do 3 miejsc po przecinku
p100 <- mapply(sym1, seq(0,1,0.1), N = 100)
p1000 <- mapply(sym1, seq(0,1,0.1), N = 1000)
```

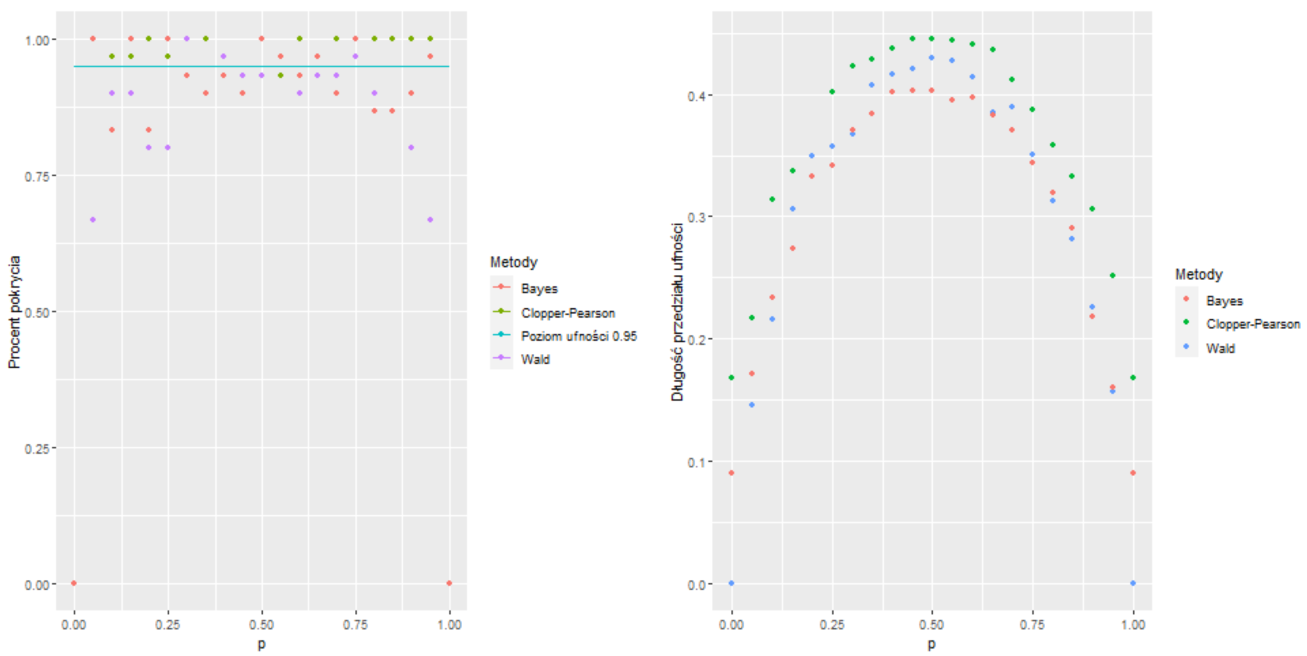
Analogicznie wyglądają funkcje do symulowania średnich długości przedziałów ufności. Kod przedstawia się następująco.

```
## pojedyncze sprawdzenie dlugosci przedzialu ufnosci
check_two <- function(n=20, p=0.6) {
    x <- binom.confint(rbinom(1, n, p), n, 0.95, "exact" )
    y <- binom.confint(rbinom(1, n, p), n, 0.95, "asymptotic" )
    z <- binom.confint(rbinom(1, n, p), n, 0.95, "bayes" )
    c(x$upper - x$lower, y$upper-y$lower, z$upper -z$lower)
}

## powtorzenie wywolania N razy w symulacji Monte Carlo dla danych p
sym2 <- function(p, N = 30, n = 20) {
    rowMeans(replicate(N, check_two(n, p)))
}

## wywolanie
d30 <- mapply(sym2, seq(0,1,0.05), N = 30)
d100 <- mapply(sym2, seq(0,1,0.05), N = 100)
d1000 <- mapply(sym2, seq(0,1,0.05), N = 1000)
```

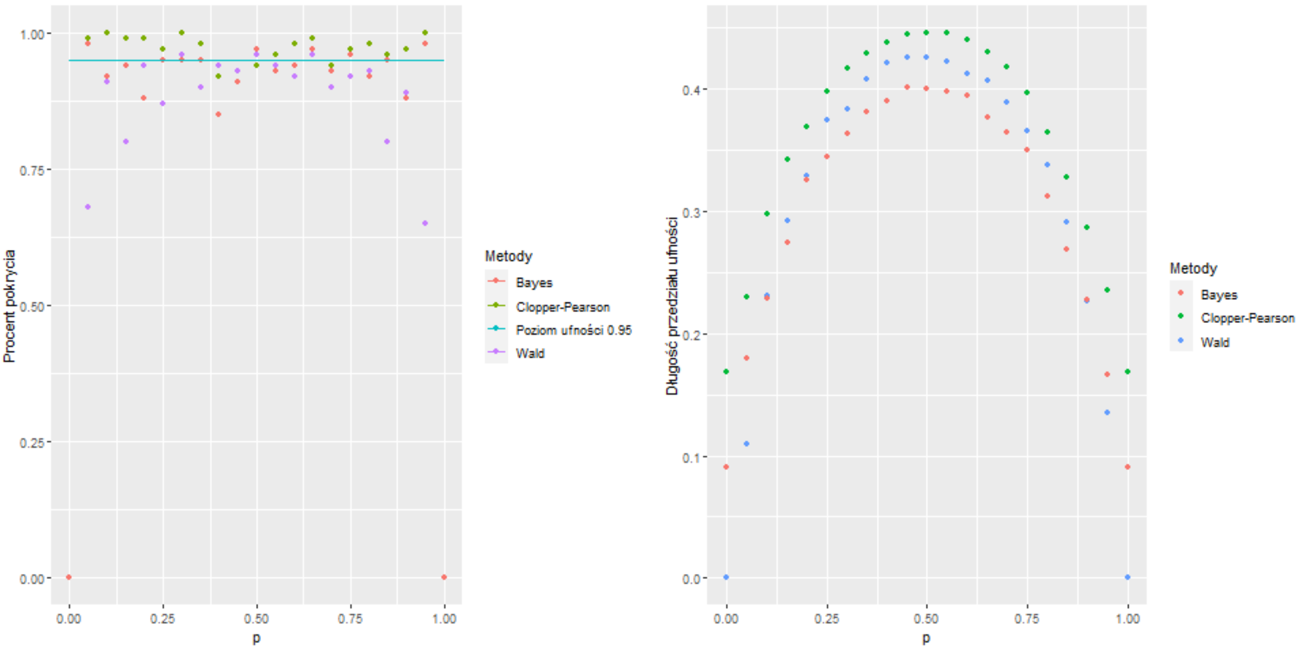
Na tej podstawie wygenerowano wykresy wartości dla kolejno $N = 30$, $N = 100$ i $N = 1000$.



(a) Procent pokrycia.

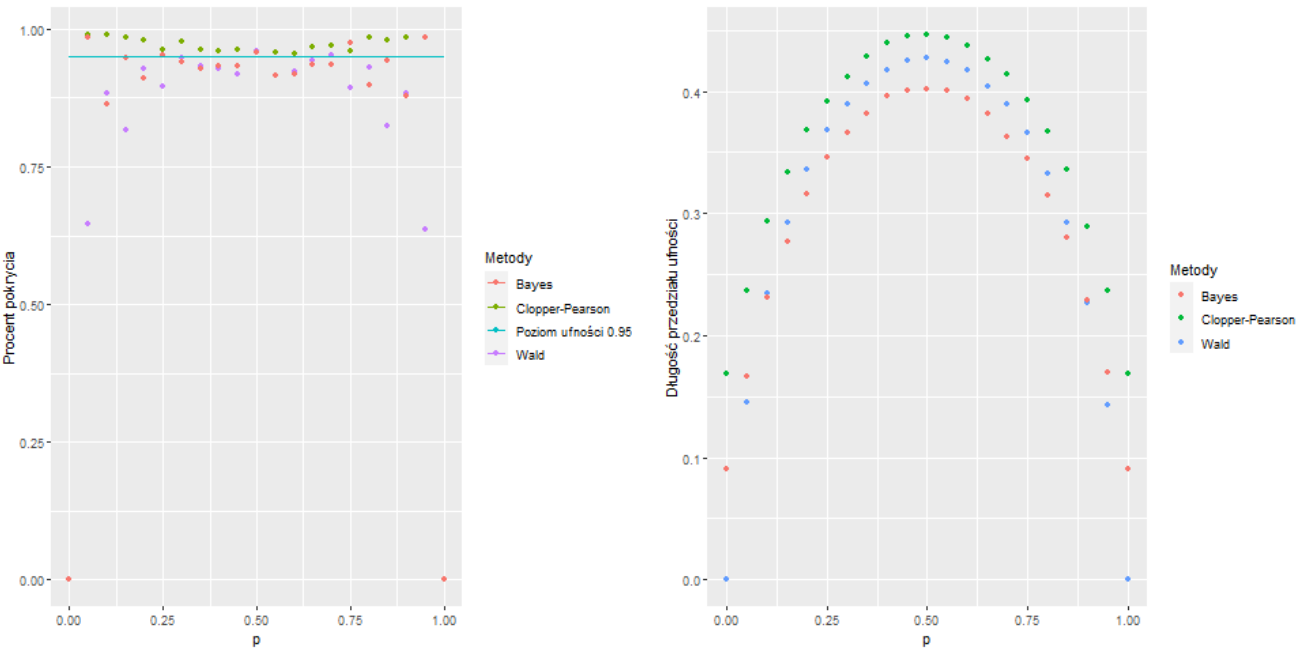
(b) Długość przedziału ufności.

Rys. 21: Procent pokrycia i długość przedziału ufności dla $N = 30$.



(a) Procent pokrycia. (b) Długość przedziału ufności.

Rys. 22: Procent pokrycia i długość przedziału ufności dla $N = 100$.



(a) Procent pokrycia. (b) Długość przedziału ufności.

Rys. 23: Procent pokrycia i długość przedziału ufności dla $N = 1000$.

Z wykresów na rysunkach 21, 22 oraz 23, ilustrujących procent pokrycia przedziałów ufności, wynika, że im większe N tym mniej rozproszone są wyniki. W każdym przypadku mniej lub bardziej oscylują one wokół wartości 0.95, czyli ustalonego poziomu ufności. Duża część wyników uzyskanych za pomocą metody Walda osiągnęła najniższe wartości, szczególnie dla granicznych wartości p bliskich 0 lub 1. Metoda Bayesa zachowuje się w podobny sposób, osiąga wartości nieco bliższe 0.95. Wyniki uzyskane dzięki metodzie Cloppera-Pearsona osiągnęły najwyższe wartości, w większości ponad poziomem ufności.

Mniejsze rozproszenie wyników wraz ze wzrostem powtórzeń N jest widoczne także na wykresach średniej długości przedziałów ufności. Tym razem różnice pomiędzy metodami są dobrze widoczne. Najdłuższe przedziały otrzymywane są za pomocą metody Cloppera Pearsona, najkrótsze - Bayesa.

Długości przedziałów mogą mieć wpływ na procent pokrycia. Do dłuższego przedziału ufności może wpaść więcej wartości. Stąd prawdopodobnie tak wysoki procent pokrycia w przypadku metody z najdłuższym średnim przedziałem ufności - metodą Cloppera-Pearsona.

Wyniki uzyskane za pomocą metod Bayesa i Walda wydają się być bardziej nieregularne, szczególnie dla granicznych wartości p. Dlatego też w tym przypadku prawdopodobniej lepiej byłoby wybrać metodę Cloppera-Pearsona. Dla $p \approx 0.25$ i $p \approx 0.75$ wszystkie metody osiągają blisko 95% pokrycia, podczas gdy metoda Bayesa ma najkrótszy przedział ufności. Może to wskazywać na pewną przewagę tej metody ponad pozostałymi w tychże przypadkach.

5. Część IV

Funkcje użyte do wykonania odpowiednich testów są przedstawione poniżej. Przypisano każdej z nich cyfrę, które będą używane jako oznaczenie odpowiedniej metody pddczas przedstawiania otrzymywanych wyników.

1. *binom.test*
2. *prop.test* z poprawką dotyczącą ciągłości (*with continuity correction*)
3. *prop.test* bez poprawki dotyczącej ciągłości (*without continuity correction*)

Przykładowe wywołania funkcji przedstawione są poniżej.

```
##binom.test z hipoteza zerowa p_0 < p
binom.test(rbinom(1, 100, 0.2), 100, p=0.5, alternative="g")

##prop.test z poprawka i hipoteza zerowa p_0 > p
prop.test(rbinom(1, 100, 0.5), 100, p=0.5, alternative="l", correct=TRUE)

##prop.test dla dwoch prob bez poprawki i z hipoteza zerowa p_1 = p_2
prop.test(c(rbinom(1, 100, 0.5), rbinom(1, 60, 0.2)), c(100, 60),
            correct=FALSE)
```

5.1. Zadanie 11.

5.1.1. Podpunkt a).

Metoda	Długość próby	Liczba sukcesów	Przedział ufności		p-wartość
	n	x	$\underline{p}$	$\overline{p}$	
1	200	71	0.289	0.426	$4.97 \cdot 10^{-5}$
2			0.290	0.426	$5.57 \cdot 10^{-5}$
3			0.292	0.423	$4.11 \cdot 10^{-5}$

Tab. 29: Testy hipotez - podpunkt a.

Test, którego wyniki przedstawiono w tabeli 29, dotyczył sprawdzenia, czy prawdopodobieństwo, że w korporacji pracuje kobieta wynosi 0.5. Długość próby w tym wypadku to ilość wszystkich pracowników, a sukcesem jest to, że pracownikiem jest kobieta. Sukcesów otrzymano 71, przez co p-wartość po zastosowaniu każdej z metod osiągnęła niską wartość, mniejszą niż ustalony poziom istotności równy 0.05. Na tej podstawie odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej. Prawdopodobieństwo, że w korporacji pracuje kobieta nie wynosi zatem 0.5.

Zauważyć także można, że otrzymane p-wartości i przedziały ufności nieco różnią się od siebie w zależności od zastosowanej metody przeprowadzania testu. Funkcją o najbardziej odstających wynikach jest *prop.test* bez poprawki dotyczącej ciągłości.

5.1.2. Podpunkt b).

Metoda	Długość próby	Liczba sukcesów	Przedział ufności		p-wartość
	n	x	$\underline{p}$	$\overline{p}$	
1	200	106	0	0.590	$< 2.2 \cdot 10^{-16}$
2			0	0.590	$< 2.2 \cdot 10^{-16}$
3			0	0.587	$< 2.2 \cdot 10^{-16}$

Tab. 30: Testy hipotez - podpunkt b.

W tym wypadku należało zbadać hipotezę zerową w postaci: prawdopodbieństwo, że pracownik jest zadowolony ze swojego wynagrodzenia jest większe bądź równe 0.8. To znaczy na 200 pracowników znaleziono 106, którzy odpowiedzieli w ankiecie pozytywnie na pytanie dotyczące ich zadowolenia z wynagrodzenia. Z tabeli 30 wynika, że p-wartości po zastosowaniu każdej z funkcji są bardzo małe, co stanowi podstawy do odrzucenia hipotezy zerowej. Prawdopodbieństwo, że pracownik jest zadowolony ze swojego wynagrodzenia jest zatem mniejsze niż 0.8

Tym razem wyniki otrzymane za pomocą *binom.test* i *prop.test* z poprawką są identyczne. Ponownie nieco inne, aczkolwiek zgodne z pozostałymi, wyniki wygenerowała metoda *prop.test* bez poprawki.

5.1.3. Podpunkt c).

Metoda	Długości prób		Liczby sukcesów		Przedział ufności		p-wartość
	n_1	n_2	x_1	x_2	$\underline{p}$	$\overline{p}$	
1	71	129	8	19	–	–	–
2					-0.141	0.072	0.6389
3					-0.130	0.061	0.4391

Tab. 31: Testy hipotez - podpunkt c.

W tym wypadku należało sprawdzić, czy prawdopodobieństwo, że kobieta pracuje na stanowisku kierowniczym jest równe prawdopodobieństwu, że mężczyzna pracuje na stanowisku kierowniczym. To jest należy porównać rozkłady prawdopodobieństwa z parametrami osobno dla obu płci na stanowiskach kierowniczych. Ilości sukcesów wyniosły odpowiednio dla kobiet 8 i mężczyzn 19. Jest to zatem sytuacja, w której należy przeprowadzić testy dla dwóch próbek jednocześnie. Funkcja *binom.test* takiej opcji nie posiada, dlatego zastosowano tylko dwie pozostałe metody.

Z tabeli 31 wynika, że obie funkcje zwróciły p-wartości większe od poziomu istotności 0.05, zatem nie ma podstaw do odrzucenia hipotezy zerowej. Rzeczywiście prawdopodobieństwo, że kobieta pracuje na stanowisku kierowniczym jest równe prawdopodobieństwu, że mężczyzna pracuje na stanowisku kierowniczym.

I tym razem pomimo zgodności ostatecznych werdyktów, otrzymane wyniki różniły się. Przedział ufności zwrócony przez metodę z poprawką jest węższy, a także p-wartość jest niższa.

5.1.4. Podpunkt d).

Metoda	Długości prób		Liczby sukcesów		Przedział ufności		p-wartość
	n_1	n_2	x_1	x_2	$\underline{p}$	$\overline{p}$	
1	71	129	36	70	–	–	–
2					-0.191	0.120	0.738
3					-0.180	0.109	0.629

Tab. 32: Testy hipotez - podpunkt d.

W niniejszym podpunkcie należało przetestować hipotezę zerową w postaci: prawdopodobieństwo, że kobieta jest zadowolona ze swojego wynagrodzenia jest równe prawdopodobieństwu, że mężczyzna jest zadowolony ze swojego wynagrodzenia. Ponownie jest to test sprawdzający równość dwóch prób, więc użyta będzie tylko funkcja *prop.test*.

Na wszystkie 71 kobiet, 36 z nich było zadowolonych ze swojego wynagrodzenia, a wśród 129 mężczyzn, 70 także odpowiedziało twierdząco na postawione pytanie. Otrzymane w tabeli 32 p-wartości są większe niż 0.05, zatem nie ma podstaw do odrzucenia hipotezy zerowej. Rzeczywiście prawdopodobieństwo, że kobieta jest zadowolona ze swojego wynagrodzenia jest równe prawdopodobieństwu, że mężczyzna jest zadowolony ze swojego wynagrodzenia.

Analogicznie, jak w poprzednim podpunkcie, przedziały ufności różnią się długością. Poprawka funkcji nieco pomniejsza $\underline{p}$ i powiększa $\overline{p}$.

5.1.5. Podpunkt e).

Metoda	Długości prób		Liczby sukcesów		Przedział ufności		p-wartość
	n_1	n_2	x_1	x_2	$\underline{p}$	$\overline{p}$	
1	71	129	23	3	–	–	–
2					-1.000	0.406	1
3					-1.000	0.395	1

Tab. 33: Testy hipotez - podpunkt e.

W tym przypadku sprawdzono, czy prawdopodobieństwo, że kobieta pracuje w dziale obsługi kadrowo-płacowej jest większe lub równe prawdopodobieństwu, że mężczyzna pracuje w dziale obsługi kadrowo-płacowej. Ponownie jest to test równości dwóch prób. Długości prób to ilości odpowiednio kobiet i mężczyzn w firmie, a sukcesem jest to, że pracują oni w dziale obsługi kadrowo-płacowej. Takich odpowiedzi było odpowiednio 23 i 3.

W tabeli 32 można odczytać, że uzyskano p-wartość równą 1, stosując obie wersje *prop.test*. Oznacza to, że należy przyjąć hipotezę zerową, a więc rzeczywiście prawdopodobieństwo, że kobieta pracuje w dziale obsługi kadrowo-płacowej jest większe lub równe prawdopodobieństwu, że mężczyzna pracuje w dziale obsługi kadrowo-płacowej.

Tym razem wartości $\underline{p}$ i $\overline{p}$ względem metod nie są tak różne. Przedział ufności dla funkcji bez poprawki jest jednak krótszy, podobnie jak w poprzednich przypadkach.