RAPORT 1
Analiza danych ankietowych

Marcelina Kosiorowska (268772)
Marcelina Białek(268871)

14.04.2024

Spis treści

Kontekst do raportu

W pewnej dużej agencji reklamowej przeprowadzono ankietę mającą na celu ocenę poziomu satysfakcji z pracy. Wzięło w niej udział dwieście losowo wybranych osób (losowanie proste ze zwracaniem). W pliku "ankieta.csv"umieszczono odpowiedzi na kilka z zadanych pytań:
• "W jakim dziale jestes zatrudniony? zmienna DZIAŁ przyjmująca wartosci:
HR (Dział obsługi kadrowo-płacowej), IT (Dział utrzymania sieci i systemów informatycznych), DK (Dział Kreatywny) lub DS (Dział Strategii),
• "Jak długo pracujesz w firmie? zmienna STAŻ przyjmująca wartości: ´ 1 (Poniżej jednego roku), 2 (Między jednym rokiem a trzema latami) lub 3 (Powyżej trzech lat),
• "Czy pracujesz na stanowisku menedżerskim? zmienna CZY_KIER przyjmująca wartości: Tak (Stanowisko menedżerskie) lub Nie (Stanowisko inne niz menedżerskie).
• "Jak bardzo zgadzasz się ze stwierdzeniem, ze firma pozwala na elastyczne godziny pracy, tym samym umozliwiając zachowanie równowagi między pracą, a życiem prywatnym?"
- zmienna PYT_1 przyjmująca wartości: ´ -2 (zdecydowanie się nie zgadzam), -1 (nie zgadzam się), 0 (nie mam zdania), 1 (zgadzam się), 2 (zdecydowanie się zgadzam).
• "Jak bardzo zgadzasz się ze stwierdzeniem, że twoje wynagrodzenie adekwatnie odzwierciedla zakres wykonywanych przez ciebie obowiązków? zmienna PYT_2 przyjmująca wartości: -2 (zdecydowanie się nie zgadzam), -1 (nie zgadzam się), 1 (zgadzam się), 2 (zdecydowanie się zgadzam).
Dodatkowo w ramach metryczki ankietowani zostali poproszeni o wskazanie swojego wieku
- zmienna WIEK przyjmująca wartosci numeryczne,
oraz wskazanie płci
- zmienna PŁEĆ przyjmująca wartość Kobieta lub Mężczyzna.
Kilka tygodni później przeprowadzono rewizję wynagrodzeń, w wyniku której część pracowników otrzymała podwyżki. Ankietowanych biorących udział w badaniu poproszono wówczas o ponowną odpowiedź na pytanie dotyczące zadowolenia z wynagrodzenia - zmienna PYT_3.

1 Zadanie 1

1.1

- Polecenie
  W pewnej duzej agencji reklamowej przeprowadzono ankietę mającą na celu ocenę poziomu satysfakcji z pracy. Wzięło w niej udział dwieście losowo wybranych osób (losowanie proste ze zwracaniem).
  Wczytaj dane i przygotuj je do analizy. Zadbaj o odpowiednie typy zmiennych, zweryfikuj czy przyjmują wartosci zgodne z powyższym opisem, zbadaj czy nie występują braki w danych.

- Kod

```
dane <- read.table(file = "ankieta.csv", sep=";", dec=",", header=TRUE)

print(head(dane))
cat('--------------------------------------------------------------\n')
braki_danych <- anyNA(dane)
if(braki_danych) {
  cat("Występują braki w danych.\n")
} else {
  cat("Brak braków danych.\n")
}
cat('--------------------------------------------------------------\n')
str(dane)
```

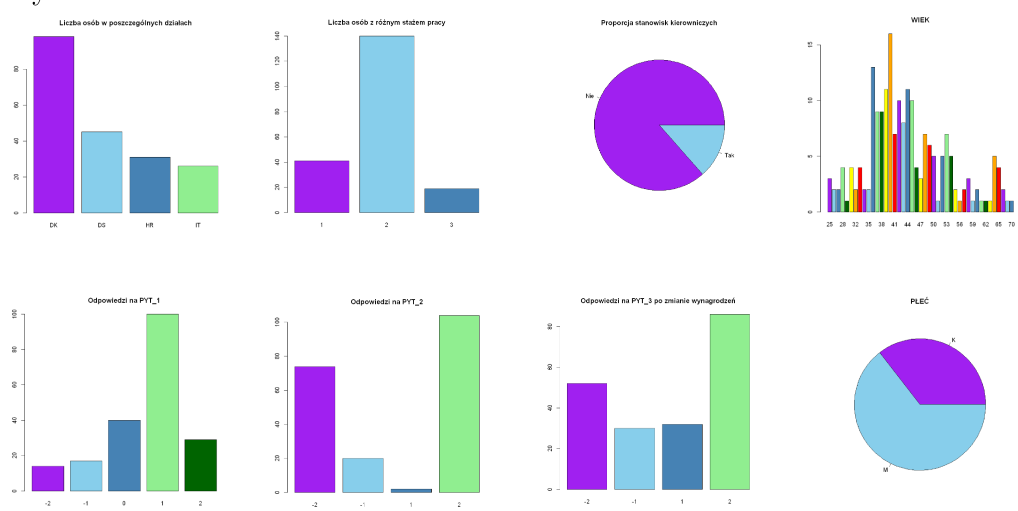
- Wyniki

```
  DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK
1    IT    2      Nie     1    -2     1    M   64
2    IT    2      Nie     0    -2    -2    M   67
3    IT    2      Nie     1     2     2    M   65
4    IT    2      Nie    -1    -2    -2    K   68
5    IT    3      Tak     1     2    -1    K   65
6    IT    3      Tak     0     1     1    K   57
--------------------------------------------------------------
Brak braków danych.
--------------------------------------------------------------
'data.frame':	200 obs. of  8 variables:
 $ DZIAŁ   : chr  "IT" "IT" "IT" "IT" ...
 $ STAŻ    : int  2 2 2 2 3 3 2 2 2 2 ...
 $ CZY_KIER: chr  "Nie" "Nie" "Nie" "Nie" ...
 $ PYT_1   : int  1 0 1 -1 1 0 2 1 1 2 ...
 $ PYT_2   : int  -2 -2 2 -2 2 1 2 -1 2 -1 ...
 $ PYT_3   : int  1 -2 2 -2 -1 1 1 -2 2 1 ...
 $ PŁEĆ    : chr  "M" "M" "M" "K" ...
 $ WIEK    : int  64 67 65 68 65 57 57 58 56 47 ...
```

- Wykresy danych
  - Kod

```
kolory <- c('purple',"skyblue", "steelblue", "lightgreen", "darkgreen","yellow", "orange", "red")

barplot(table(dane$DZIAŁ), main="Liczba osób w poszczególnych działach", col=kolory)
barplot(table(dane$STAŻ), main="Liczba osób z różnym stażem pracy", col=kolory)
barplot(table(dane$CZY_KIER), main="Proporcja stanowisk kierowniczych", col=kolory)
barplot(table(dane$PYT_3), main="Odpowiedzi na PYT_3 po zmianie wynagrodzeń", col=kolory)
barplot(table(dane$WIEK), main="WIEK", col=kolory)


barplot(table(dane$PYT_1), main="Odpowiedzi na PYT_1", col=kolory)
barplot(table(dane$PYT_2), main="Odpowiedzi na PYT_2", col=kolory)
pie(table(dane$PŁEĆ), main="PŁEĆ", col=kolory)
```

  - Wyniki



1.2

- Polecenie
  Utwórz zmienną WIEK_KAT, przeprowadzając kategoryzację zmiennej WIEK korzystając z następujących przedziałów: do 35 lat, między 36 a 45 lat, między 46 a 55 lat, powyżej 55 lat.

- Kod

```r
dane$WIEK_KAT <- cut(dane$WIEK, breaks = c(0, 35, 45, 55, Inf),
                     labels = c("do 35 lat", "między 36 a 45 lat", "między 46 a 55 lat", "powyżej 55 lat"),
                     include.lowest = TRUE)

head(dane)
```

- Wynik

A data.frame: 6 × 9

	DZIAŁ	STAŻ	CZY_KIER	PYT_1	PYT_2	PYT_3	PŁEĆ	WIEK	WIEK_KAT
	<chr>	<int>	<chr>	<int>	<int>	<int>	<chr>	<int>	<fct>
1	IT	2	Nie	1	-2	1	M	64	powyżej 55 lat
2	IT	2	Nie	0	-2	-2	M	67	powyżej 55 lat
3	IT	2	Nie	1	2	2	M	65	powyżej 55 lat
4	IT	2	Nie	-1	-2	-2	K	68	powyżej 55 lat
5	IT	3	Tak	1	2	-1	K	65	powyżej 55 lat
6	IT	3	Tak	0	1	1	K	57	powyżej 55 lat

1.3

- Polecenie
  Sporządź tablice licznosci dla zmiennych: DZIAŁ, STAŻ, CZY_KIER, PŁEĆ, WIEK_KAT.

- Kod

```
licznosc_dzial = table(dane$DZIAŁ)
licznosc_staz = table(dane$STAŻ)
licznosc_czy_kier = table(dane$CZY_KIER)
licznosc_plec = table(dane$PŁEĆ)
licznosc_wiek_kat = table(dane$WIEK_KAT)

df_licznosc_dzial_t = as.data.frame(t(df_licznosc_dzial))
df_licznosc_staz_t = as.data.frame(t(df_licznosc_staz))
df_licznosc_czy_kier_t = as.data.frame(t(df_licznosc_czy_kier))
df_licznosc_plec_t = as.data.frame(t(df_licznosc_plec))
df_licznosc_wiek_kat_t = as.data.frame(t(df_licznosc_wiek_kat))


df_licznosc_dzial_t
df_licznosc_staz_t
df_licznosc_czy_kier_t
df_licznosc_plec_t
df_licznosc_wiek_kat_t
```

- Wynik

A data.frame: 2 × 4

	V1	V2	V3	V4
	<chr>	<chr>	<chr>	<chr>
Var1	DK	DS	HR	IT
Freq	98	45	31	26

DZIAŁ

A data.frame: 2 × 3

	V1	V2	V3
	<chr>	<chr>	<chr>
Var1	1	2	3
Freq	41	140	19

STAŻ

A data.frame: 2 × 2

	V1	V2
	<chr>	<chr>
Var1	Nie	Tak
Freq	173	27

CZY_KIER

A data.frame: 2 × 2

	V1	V2
	<chr>	<chr>
Var1	K	M
Freq	71	129

PŁEĆ

A data.frame: 2 × 4

	V1	V2	V3	V4
	<chr>	<chr>	<chr>	<chr>
Var1	do 35 lat	między 36 a 45 lat	między 46 a 55 lat	powyżej 55 lat
Freq	26	104	45	25

WIEK_KAT

Tablice liczności dla wybranych zmiennych ze zbioru danych

1.4

- Polecenie
  Sporządź wykresy kołowe oraz wykresy słupkowe dla zmiennych: PYT_1 oraz PYT_2

- Kody

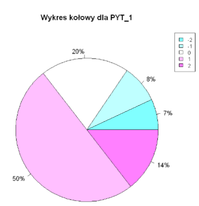
```r
tab_PYT_1 <- table(dane$PYT_1)
procenty_PYT_1 <- round(prop.table(tab_PYT_1) * 100)
lbls_PYT_1 <- paste(procenty_PYT_1, "%", sep="")
kolory_PYT_1 <- cm.colors(length(tab_PYT_1))
pie(tab_PYT_1, labels = lbls_PYT_1, col = kolory_PYT_1, main="Wykres kołowy dla PYT_1")
legend("topright", legend = names(tab_PYT_1), fill = kolory_PYT_1, cex = 0.8)


tab_PYT_2 <- table(dane$PYT_2)
procenty_PYT_2 <- round(prop.table(tab_PYT_2) * 100)
lbls_PYT_2 <- paste(procenty_PYT_2, "%", sep="")
kolory_PYT_2 <- topo.colors(length(tab_PYT_2))
pie(tab_PYT_2, labels = lbls_PYT_2, col = kolory_PYT_2, main="Wykres kołowy dla PYT_2")
legend("topright", legend = names(tab_PYT_2), fill = kolory_PYT_2, cex = 0.8)
```

Kod do wykresów kołowych

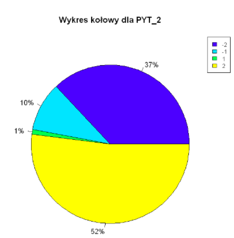
```r
barplot(table(dane$PYT_1), main="Wykres słupkowy dla zmiennej PYT_1", xlab="Odpowiedzi", ylab="Liczba odpowiedzi",
        col=cm.colors(length(table(dane$PYT_1))))
barplot(table(dane$PYT_2), main="Wykres słupkowy dla zmiennej PYT_2", xlab="Odpowiedzi", ylab="Liczba odpowiedzi",
        col=topo.colors(length(table(dane$PYT_1))))
```

Kod do wykresów słupkowych

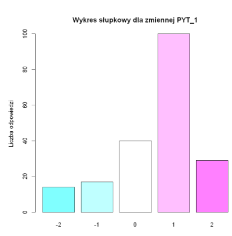
- Wyniki



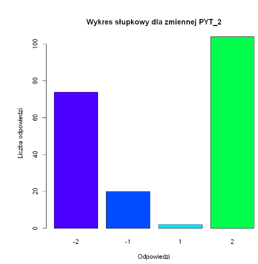
Wykres kołowy dla zmiennej PYT_1



Wykres kołowy dla zmiennej PYT_2



Wykres słupkowy dla zmiennej PYT_1



Wykres słupkowy dla zmiennej PYT_2

1.5

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

- Kod

```
ftable(dane$PYT_1, dane$DZIAŁ)
ftable(dane$PYT_1, dane$STAŻ)
ftable(dane$PYT_1, dane$CZY_KIER)
ftable(dane$PYT_1, dane$PŁEĆ)
ftable(dane$PYT_1, dane$WIEK_KAT)
```

- Wyniki

	DK	DS	HR	IT
-2	9	3	2	0
-1	10	3	2	2
0	17	14	5	4
1	51	15	19	15
2	11	10	3	5

PYT_1 i DZIAŁ

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

PYT_1 i STAŻ

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

PYT_1 i CZY_KIER

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

PYT_1 i PŁEĆ

	do 35 lat	między 36 a 45 lat	między 46 a 55 lat	powyżej 55 lat
-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

PYT_1 i WIEK_KAT

Tablice wielodzielcze dla wybranych par zmiennych ze zbioru danych

1.6

- Polecenie
  Sporządź tablicę wielodzielczą dla pary zmiennych: PYT_2 i PYT_3.
- Kod

```
ftable(dane$PYT_2, dane$PYT_3)
```

- Wynik

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

Tablica wielodzielcza dla pary zmiennych PYT_2 i PYT_3

1.7

- Polecenie
  Utwórz zmienną CZY_ZADOW na podstawie zmiennej PYT_2 łacząc kategorie "nie zgadzam się"i "zdecydowanie się nie zgadzam"oraz "zgadzam się"i "zdecydowanie się zgadzam".
- Przyjęłyśmy wersję utworzenia zmiennej w oparciu o zmienne, gdzie wartości -2 i -1 odpowiadają niezadowoleniu, a wartości 1 i 2 zadowoleniu.
- Kod

```
dane$CZY_ZADOW <- ifelse(dane$PYT_2 %in% c(-2, -1), "niezadowolony",
                         ifelse(dane$PYT_2 %in% c(1, 2), "zadowolony", NA))
head(dane)
```

- Wynik

A data.frame: 6 × 10

	DZIAŁ	STAŻ	CZY_KIER	PYT_1	PYT_2	PYT_3	PŁEĆ	WIEK	WIEK_KAT	CZY_ZADOW
	<chr>	<int>	<chr>	<int>	<int>	<int>	<chr>	<int>	<fct>	<chr>
1	IT	2	Nie	1	-2	1	M	64	powyżej 55 lat	niezadowolony
2	IT	2	Nie	0	-2	-2	M	67	powyżej 55 lat	niezadowolony
3	IT	2	Nie	1	2	2	M	65	powyżej 55 lat	zadowolony
4	IT	2	Nie	-1	-2	-2	K	68	powyżej 55 lat	niezadowolony
5	IT	3	Tak	1	2	-1	K	65	powyżej 55 lat	zadowolony
6	IT	3	Tak	0	1	1	K	57	powyżej 55 lat	zadowolony

Dane z utworzoną zmienną CZY_ZADOW

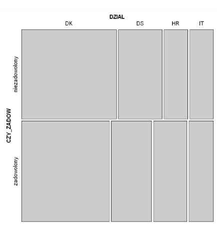
1.8

- Polecenie
  Korzystając z funkcji mosaic z biblioteki vcd, sporządź wykresy mozaikowe odpowiadające parom zmiennych: CZY_ZADOW i DZIAŁ, CZY_ZADOW i STAŻ, CZY_ZADOW i CZY_KIER, CZY_ZADOW i PŁEC´ oraz CZY_ZADOW i WIEK_KAT. Czy na podstawie uzyskanch wykresów mozna postawić pewne hipotezy dotyczące relacji między powyższymi zmiennymi? Spróbuj sformułować kilka takich hipotez.

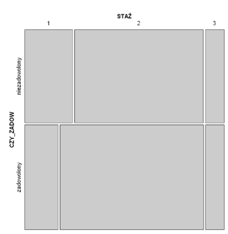
- Kod

```
mosaic(~CZY_ZADOW + DZIAŁ, dane)
mosaic(~CZY_ZADOW + STAŻ, dane)
mosaic(~CZY_ZADOW + CZY_KIER, dane)
mosaic(~CZY_ZADOW + PŁEĆ, dane)
mosaic(~CZY_ZADOW + WIEK_KAT, dane)
```

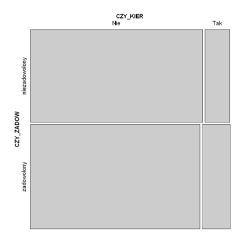
- Wyniki



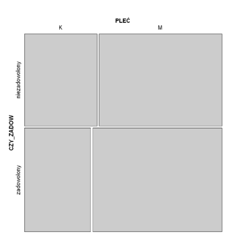
PYT_1 i DZIAŁ



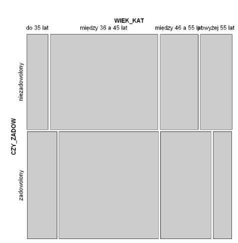
PYT_1 i STAŻ



PYT_1 i CZY_KIER



PYT_1 i PŁEĆ



PYT_1 i WIEK_KAT

Tablice mozaikowe dla wybranych par zmiennych ze zbioru danych

- Hipotezy

  - Istnieje zależność między działem, w którym pracownicy są zatrudnieni, a ich postrzeganiem wysokości wynagrodzenia. Na przykład, większość pracowników działu kreatywnego (DK) postrzegają wynagrodzenie jako niezadowalające, a pracownicy działu kadr (HR) jako zadowalające. Prawdopodobnie jest to powiązane z wysokością wynagrodzenia w stosunku do nakładu pracy oraz ilości obowiązków.
  - Zadowolenie pracowników może być powiązane z długością ich stażu w firmie. Na przykład, pracownicy z najkrótszym stażem mogą być bardziej niezadowoleni z wypłaty z powodu braku przyzwyczajenia do pracy lub większych ambicji zarobkowych.
  - Bycie na stanowisku kierowniczym może być powiązane zarówno wyższym, jak i niższym poziomem zadowolenia z wypłaty od zwykłych pracowników, ponieważ prawdopodobnie wyższą wypłatę mogą postrzegać jako adekwatną lub nieadekwatną do większej liczby obowiązków oraz odpowiedzialności.
  - To, że większość kobiet deklaruje niezadowolenie z wynagrodzenia może być spowodowane różnicami w równości wynagrodzeń lub innymi aspektami równości płci w miejscu pracy.
  - Zadowolenie pracowników z wypłaty różniące się od grupy wiekowej może być spowodowane różnymi postawami w zależności od etapu życia. To, że większość pracowników w wieku powyżej 55 lat jest niezadowolonych, może być spowodowane zarówno wypaleniem zawodowym, jak i nieadekwatnością wynagrodzenia w stosounku do doświadczenia.

2 Zadanie 2

- Polecenie
  Zapoznaj się z funkcjami summary oraz plot (wykresy typu ”bar”, ”heat”oraz ”density”), a nastepnie zilustruj odpowiedzi na pytanie: ”Jak bardzo zgadzasz sie ze stwierdzeniem, ze firma pozwala na (...)?”(zmienna ˙ PYT_1) w całej badanej grupie oraz w podgrupach ze wzgledu na zmienną CZY_KIER.

- Kod

```
df <- dane
df$PYT_1 <- factor(df$PYT_1, levels = c(-2, -1, 0, 1, 2),
labels = c("zdecydowanie się nie zgadzam", "nie zgadzam się", "nie mam zdania", "zgadzam się", "zdecydowanie się zgadzam"))

df$CZY_KIER <- factor(df$CZY_KIER, levels = c("Tak","Nie"))

# a) działanie funkcji summary

# w całej grupie

likert_pyt_1 <- likert(df[,"PYT_1", drop=FALSE])
summary_pyt_1 <- summary(likert_pyt_1)
print(summary_pyt_1)

# ze względu na podgrupę kierowników i niekierowników

likert_pyt_1_czy_kier <- likert(df[, "PYT_1", drop = FALSE], grouping = df$CZY_KIER)
summary_pyt_1_czy_kier <- summary(likert_pyt_1_czy_kier)
print(summary_pyt_1_czy_kier)

# b) plots

# wykresy typu "bar", "heat" i "density" bez grupowania
plot(likert_pyt_1, type = "bar", centered = FALSE)
plot(likert_pyt_1, type = "heat", centered = FALSE)
plot(likert_pyt_1, type = "density")

# wykresy typu "bar" z grupowaniem
plot(likert_pyt_1_czy_kier, type = "bar", centered = FALSE)
```

- Wyniki
  - Działanie funkcji summary

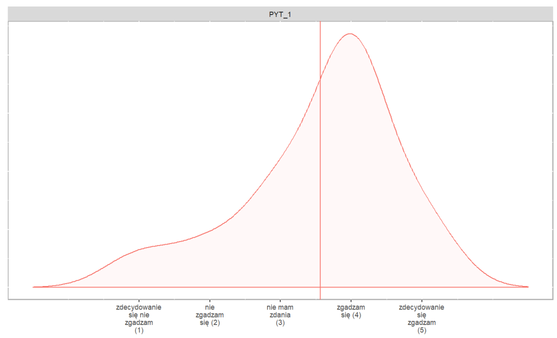
	Item	low	neutral	high	mean	sd
1	PYT_1	15.5	20	64.5	3.565	1.063688

Wyniki funkcji summary dla całej grupy

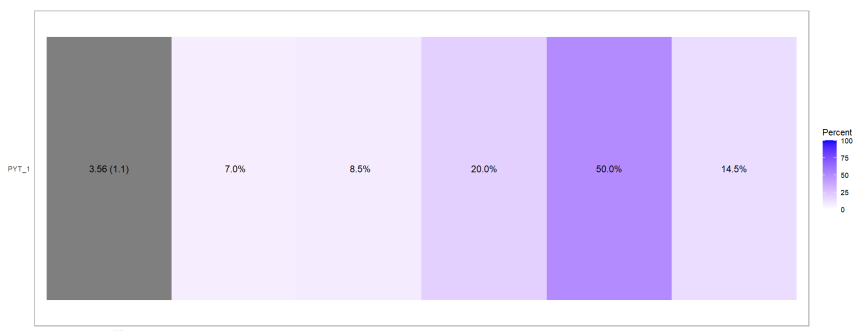
	Group	Item	low	neutral	high	mean	sd
1	Nie	PYT_1	13.87283	19.65318	66.47399	3.624277	1.030291
2	Tak	PYT_1	25.92593	22.22222	51.85185	3.185185	1.210119

Wyniki funkcji summary z podziałem ze względu na podgrupy CZY_KIER

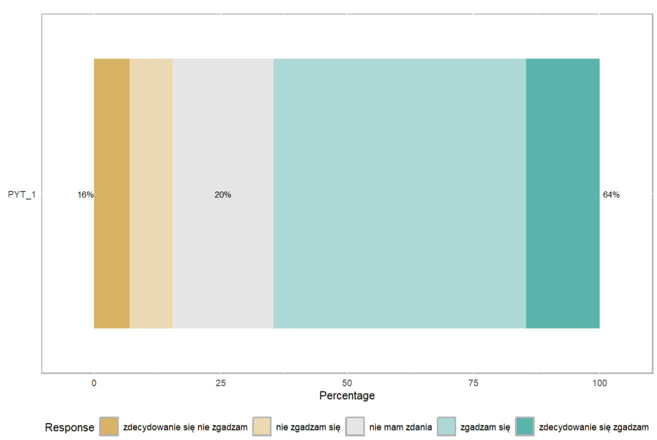
- Działanie funkcji plot



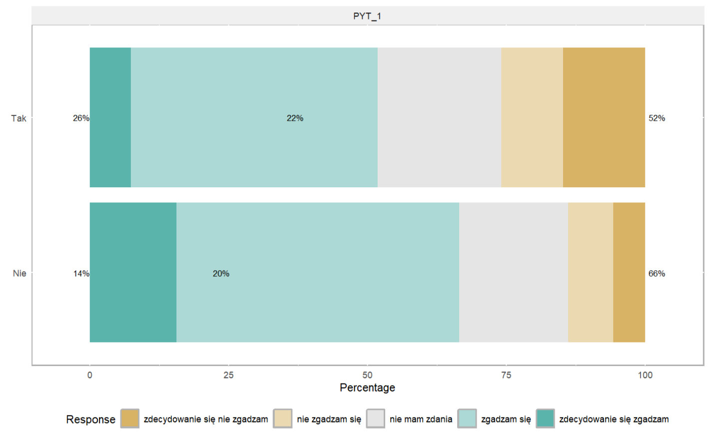
Wykres gęstości rozkładu odpowiedzi na PYT_1 w całej grupie



Wykres heat rozkładu odpowiedzi na PYT_1 w całej grupie



Wykres bar rozkładu odpowiedzi na PYT_1 w całej grupie



Wykres bar rozkładu odpowiedzi na PYT_1 w podgrupie CZY_KIER

- Komentarz
  Odczytujemy, że w badanej grupie zdecydowana większość zgadza się z tezą stawianą w PYT_1 i stanowi ona 64.5% grupy (z wykresu heat i funkcji summary). Dodatkowo zaobserwowałyśmy, iż wśród kierowników stosunek do PYT_1 jest odmienny, kierownicy są sceptyczniej nastawieni do tezy tego pytania.

3 Zadanie 3

- Polecenie
  Zapoznaj sie z funkcją sample z biblioteki stats, a nastepnie wylosuj próbke o licznosci 10% wszystkich rekordów z pliku "ankieta.csv"w dwóch wersjach: ze zwracaniem oraz bez zwracania.

- Kod

```
# bez zwracania
ind <- sample(nrow(dane), 1/10 * nrow(dane), replace = FALSE)
print(dane[ind,])

# ze zwracaniem
ind2 <- sample(nrow(dane), 1/10 * nrow(dane), replace = TRUE)
print(dane[ind2,])
```

- Wyniki

	DZIAŁ	STAŻ	CZY_KIER	PYT_1	PYT_2	PYT_3	PŁEĆ	WIEK
175	HR	2	Nie	1	2	1	M	36
68	DK	2	Nie	1	2	2	M	59
58	DK	2	Nie	1	2	2	M	47
29	DK	1	Nie	1	2	2	M	26
79	DK	1	Nie	1	-1	2	K	39
32	DK	1	Nie	1	2	2	M	32
168	DS	2	Nie	0	-2	-2	M	58
183	HR	2	Nie	1	2	2	M	40
124	DK	2	Nie	2	-2	-2	M	42
107	DK	2	Nie	1	2	2	K	36
75	DK	2	Nie	2	2	2	K	44
70	DK	1	Nie	-1	-2	-1	M	25
109	DK	2	Tak	-2	-2	-2	K	40
120	DK	2	Nie	0	-2	-2	M	39
190	HR	2	Nie	1	2	2	M	49
74	DK	2	Nie	1	2	1	M	33
85	DK	1	Nie	1	2	2	M	54
178	HR	3	Tak	0	-2	1	M	40
9	IT	2	Tak	1	-1	-2	K	58
137	DS	2	Nie	-2	-2	2	K	42

	DZIAŁ	STAŻ	CZY_KIER	PYT_1	PYT_2	PYT_3	PŁEĆ	WIEK
115	DK	1	Nie	-1	-2	-2	M	44
122	DK	2	Tak	1	-1	1	M	41
71	DK	2	Tak	-1	2	2	M	34
52	DK	1	Nie	-2	-2	-1	M	43
60	DK	2	Nie	1	-2	-2	M	49
54	DK	1	Nie	1	2	1	M	40
94	DK	2	Nie	2	2	2	M	37
94.1	DK	2	Nie	2	2	2	M	37
15	IT	2	Nie	1	2	2	K	53
85	DK	1	Nie	1	2	2	M	54
52.1	DK	1	Nie	-2	-2	-1	M	43
142	DS	2	Nie	1	2	2	K	45
3	IT	2	Nie	0	-2	-2	M	67
130	DS	2	Nie	-1	-2	-2	K	36
135	DS	2	Nie	2	2	2	K	36
146	DS	3	Nie	0	-2	-2	K	46
72	DK	2	Nie	1	2	1	M	28
27	IT	3	Nie	1	2	2	K	51
126	DS	2	Nie	1	2	2	K	40
26	IT	2	Nie	1	-1	-1	K	50

Rysunek 6: Próbka danych o rozmiarze 10% z całej grupy bez zwracania (po lewej) oraz ze zwracaniem (po prawej)

4 Zadanie 4

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

- Opis metody
  Używamy rozkładu jednostajnego do generowania liczb losowych z przedziału od 0 do 1. Następnie liczbę wygenerowaną z rozkładu jednostajnego porównujemy z wartością p. Jeśli wygenerowana liczba jest mniejsza niż p, oznacza to sukces (1), w przeciwnym razie jest to porażka (0). Po przeprowadzeniu n prób i określeniu, które z nich są sukcesami, sumujemy liczbę sukcesów. Ta suma to wartość zmiennej z rozkładu dwumianowego dla jednej próbki.

- Kod

```r
generate_binomial <- function(n, p, size) {
  samples <- numeric(size)
  for (i in 1:size) {
    #generowanie zmiennych z rozkładu jednostajnego
    trials <- ifelse(runif(n) < p, 1, 0)
    samples[i] <- sum(trials)
  }
  return(samples)
}

set.seed(123) #ziarno
n <- 500
p_wart <- c(0.1, 0.5, 0.9)
rozmiar <- 1000

for (p in p_wart) {
  samples <- generate_binomial(n, p, rozmiar)

  #porównanie empirycznego i teoretycznego p
  p_empiryczne <- mean(samples) / n
  cat(sprintf("Dla p = %s:\n", p))
  cat(sprintf("Empiryczne p: %f\n", p_empiryczne))
  cat(sprintf("Teoretyczne p: %f\n\n", p))

  #porównanie dystrybuant
  ecdf_fun <- ecdf(samples)
  plot(ecdf_fun, main = paste("Porównanie dystrybuant dla p =", p),
       xlab = "Wartości", ylab = "Dystrybuanta", col = "blue", lwd = 2)
  x <- seq(0, n, length.out = 1000)
  y <- pbinom(x, n, p)
  lines(x, y, col = "red", lwd = 2)
  legend("bottomright", legend = c("Teoretyczna", "Empiryczna"),
         col = c("red", "blue"), lwd = 2, bty = "n")
}
```

- Wyniki

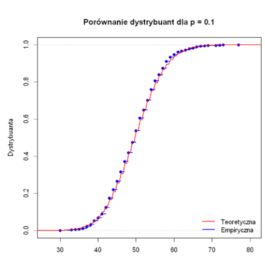
  - Porównanie wartości p dla teoretycznych i empirycznych wartości

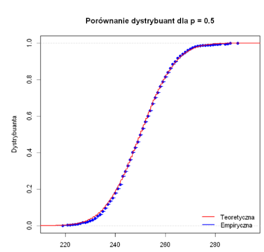
```
Dla p = 0.1:
Empiryczne p: 0.099910
Teoretyczne p: 0.100000
```

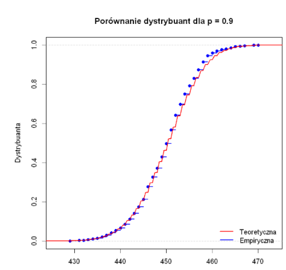
```
Dla p = 0.5:
Empiryczne p: 0.501356
Teoretyczne p: 0.500000
```

```
Dla p = 0.9:
Empiryczne p: 0.900496
Teoretyczne p: 0.900000
```

  - Porównanie dystrybuant teoretycznych i empirycznych dla każdej z analizowanych wartości p







5 Zadanie 5

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

- Opis metody
  Najpierw używamy funkcji sample, aby wybrać jednen z możliwych wyników w każdej próbie, opierając się na podanych prawdopodobieństwach. Następnie sumujemy liczbę sukcesów w poszczególnych kategoriach. Zwracamy wektor z liczbą zdarzeń dla każdego wyniku.

- Kod

```r
#funkcja do generowania realizacji z rozkładu wielomianowego
generate_multinomial <- function(n, prob) {
  num_classes <- length(prob)
  samples <- numeric(num_classes)
  for (i in 1:n) {
    class <- sample(1:num_classes, 1, prob = prob)
    samples[class] <- samples[class] + 1
  }

  return(samples)
}

p_values <- list(c(0.1, 0.3, 0.4, 0.2), c(0.5, 0.3, 0.2), c(0.1, 0.1, 0.3, 0.2, 0.3))
n_values <- c(1000, 1000, 1000)

for (i in 1:length(p_values)) {
  p <- p_values[[i]]
  n <- n_values[i]
  cat("Parametry p:", p, "\n")
  cat("Liczba prób n:", n, "\n")
  samples <- generate_multinomial(n, p)

  empirical_prob <- samples / sum(samples)
  cat("Empiryczne prawdopodobieństwa:", empirical_prob, "\n")

  cat("Teoretyczne prawdopodobieństwa:", p, "\n\n")

  barplot(rbind(empirical_prob, p), beside = TRUE,
          col = c("green", "orange"),
          legend.text = c("Empiryczne", "Teoretyczne"),
          main = paste("p =", paste(p, collapse = ", ")),
          xlab = "", ylab = "Prawdopodobieństwo")
}
```

- Wyniki

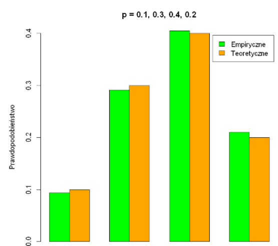
  - Porównanie wartości p dla teoretycznych i empirycznych wartości

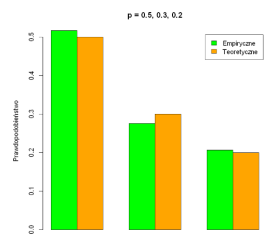
```
Parametry p: 0.1 0.3 0.4 0.2
Liczba prób n: 1000
Empiryczne prawdopodobieństwa: 0.108 0.276 0.414 0.202
Teoretyczne prawdopodobieństwa: 0.1 0.3 0.4 0.2
```

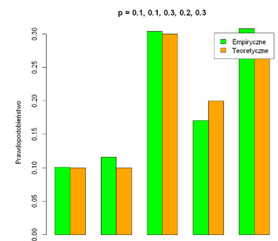
```
Parametry p: 0.5 0.3 0.2
Liczba prób n: 1000
Empiryczne prawdopodobieństwa: 0.513 0.298 0.189
Teoretyczne prawdopodobieństwa: 0.5 0.3 0.2
```

```
Parametry p: 0.1 0.1 0.3 0.2 0.3
Liczba prób n: 1000
Empiryczne prawdopodobieństwa: 0.091 0.112 0.297 0.197 0.303
Teoretyczne prawdopodobieństwa: 0.1 0.1 0.3 0.2 0.3
```

  - Porównanie wartości p dla teoretycznych i empirycznych wartości na wykresach słupkowych







6 Zadanie 6

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

- Kod

```r
clopper_pearson <- function(p_ufnosci, args_list){
  if (is.vector(args_list[[2]])){  # jeśli pojawi się wektor to obsługujemy 1 przypadek
    dane <- args_list[[2]] # pobranie wektora danych
    s <- sum(dane == "TAK") # liczba sukcesów
    n <- length(dane) # liczba wszystkich prób

  } else if (length(args_list) == 3) {  # trzy elementy oznaczają 2 przypadek
    s <- args_list[[2]] # liczba sukcesów, podanych na sztywno
    n <- args_list[[3]] # liczba prób

  } else { # ewentualne wyeliminowanie błędów
    return("Niepoprawna liczba argumentów.")
  }

  # wyznaczenie przedziału ufności
  start_przedzial <- qbeta(1/2-p_ufnosci/2, s, n-s+1)
  koniec_przedzial <- qbeta(1/2+p_ufnosci/2, s+1, n-s)

  return(paste("[", start_przedzial, ", ", koniec_przedzial, "]", sep = ""))
}
```

Funkcja wyznaczająca p.ufności Cloppera-Pearsona z komentarzem

- Przykładowe działanie

```
> print(clopper_pearson(95/100, list(1, 900)))
[1] "[0, 0.975]"
> print(clopper_pearson(95/100, c("NIE", "TAK", "TAK")))
[1] "[0.025, 1]"
>
```

Przedziały ufności wyznaczone dla 2 przypadków: poprzez poziom ufności, liczbę sukcesów, liczbę prób ORAZ przez poziom ufnosci i wektor danych

7 Zadanie 7

- Polecenie
  Korzystając z funkcji napisanej w zadaniu 6. wyznacz realizacje przedziałów ufnosci dla prawdopodobieństwa, że pracownik jest zadowolony z wynagrodzenia w pierwszym ˙ badanym okresie oraz w drugim badanym okresie. Skorzystaj ze zmiennych CZY_ZADW oraz CZY_ZADW_2 (utwórz zmienne analogicznie jak w zadaniu 1.7). Przyjmij $1 - \alpha = 0.95$

- Kod

```
dane$CZY_ZADW <- ifelse(dane$PYT_2 %in% c(1, 2), "TAK", ifelse(dane$PYT_2 %in% c(-1, -2), "NIE", NA))

dane$CZY_ZADW_2 <- ifelse(dane$PYT_3 %in% c(1, 2), "TAK", ifelse(dane$PYT_3 %in% c(-1, -2), "NIE", NA))

# Wyświetlanie wyników dla CZY_ZADW
cat("Przedział ufności (zadowolenie z wynagrodzenia w 1 okresie):\n")
print(clopper_pearson(0.95, dane$CZY_ZADW))

# Wyświetlanie wyników dla CZY_ZADW_2
cat("Przedział ufności (zadowolenie z wynagrodzenia w 2 okresie):\n")
print(clopper_pearson(0.95, dane$CZY_ZADW_2))
```

- Wyniki

```
Przedział ufności (zadowolenie z wynagrodzenia w 1 okresie):
> print(clopper_pearson(0.95, dane$CZY_ZADW))
[1] "[0, 0.975]"
> # Wyświetlanie wyników dla CZY_ZADW_2
> cat("Przedział ufności (zadowolenie z wynagrodzenia w 2 okresie):\n")
Przedział ufności (zadowolenie z wynagrodzenia w 2 okresie):
> print(clopper_pearson(0.95, dane$CZY_ZADW_2))
[1] "[0, 0.975]"
>
```

Przedziały ufnosci dla prawdopodobieństwa, że pracownik jest zadowolony z wynagrodzenia w 1 badanym okresie oraz w 2 badanym okresie.

8 Zadanie 9

- Polecenie
  Przeprowadź symulacje, których celem jest porównanie prawdopodobienstwa pokrycia i długosci przedziałów ufności Cloppera-Pearsona, Walda i trzeciego dowolnego typu zaimplementowanego w funkcji binom.confint. Rozważ $1\alpha = 0.95$, rozmiar próby $n \in 30, 100, 1000$ i rózne wartości prawdopodobieństwa p. Wyniki umieść na wykresach i sformułuj wnioski, które dla konkretnych danych ułatwią wybór konkretenego typu przedziału ufności.

- Opis kodu
  Na początku symulacji określamy wartości p oraz trzy różne wielkości n zadane w wymaganiach. Będziemy badać trzy metody obliczania przedziałów ufności: dokładnej (Cloppera-Pearsona), asymptotycznej (Walda) i wybranej przez nas metody Wilsona. Dla każdej kombinacji wartości p, wielkości próby i metody obliczania przedziałów ufności, wykonujemy 100 symulacji. W każdej symulacji losowo generujemy próbę. Następnie dla każdej symulacji sprawdzamy, czy prawdziwa wartość p znajduje się w obliczonym przedziale ufności. To daje nam informację o pokryciu oraz obliczamy długość tego przedziału. Potem podsumowujemy wyniki, obliczając średnie pokrycie i średnią długość przedziałów ufności dla każdej metody i konkretnej wartości parametrów. Na końcu wyświetlamy wyniki na wykresach.

- Kody

```r
p_values <- seq(0.01, 0.99, by = 0.01)
n_values <- c(30, 100, 1000)
MC <- 100
methods <- c("exact", "asymptotic", "wilson")
DF <- data.frame(Method = character(), N = numeric(), P = numeric(), Coverage = numeric(), Length = numeric(), stringsAsFactors = FALSE)

for (i in p_values) {
    for (n in n_values) {
        for (method in methods) {
            coverage <- numeric(MC)
            dlugosc_przedzialu_ufnosci <- numeric(MC)
            for (step in 1:MC) {
                x <- rbinom(1, n, i)
                przedzial_ufnosci <- binom.confint(x, n = n, conf.level = 0.95, method = method)
                sprawdz <- (przedzial_ufnosci$lower <= i) & (i <= przedzial_ufnosci$upper)
                if (any(sprawdz)) {
                    dlugosc_przedzialu_ufnosci[step] <- przedzial_ufnosci$upper[sprawdz] - przedzial_ufnosci$lower[sprawdz]
                    coverage[step] <- as.numeric(sprawdz)
                }
            }
            df <- data.frame(Method = method, N = n, P = i, Coverage = sum(coverage) / MC *100, Length = mean(dlugosc_przedzialu_ufnosci), stringsAsFacto
            DF <- rbind(DF, df)
        }
    }
}

filtered_DF <- DF[DF$P %in% seq(0.01, 0.99, by = 0.20), ]
filtered_DF_list <- split(filtered_DF, filtered_DF$N)
for (df in filtered_DF_list) {
  print(df)
}
```

Główna pętla kodu

```r
for (n_val in unique(DF$N)) {
  df_subset <- subset(DF, N == n_val)

  #wykres dla procentu pokrycia
  p_coverage <- ggplot(df_subset, aes(x = P, y = Coverage, group = Method, color = Method)) +
    geom_line(linewidth = 1.2) +
    geom_hline(yintercept = 0.95, linetype = "dashed") +
    geom_hline(yintercept = 0.5, linetype = "dashed", color = "NA") +
    geom_vline(xintercept = c(0.3, 0.5, 0.8), linetype = "dashed") +
    scale_x_continuous(breaks = seq(0, 1, 0.1)) +
    labs(x = "Wartość prawdopodobieństwa p", y = "Procent pokrycia") +
    annotate("text", x = 0.08, y = 0.955, label = "Poziom uf 0.95", size = 3) +
    facet_wrap(~ N) +
    ggtitle(paste("Procent pokrycia (n =", n_val, ")"))

  print(p_coverage)

  #wykres dla długości przedziału ufności
  p_length <- ggplot(df_subset, aes(x = P, y = Length, group = Method, color = Method)) +
    geom_line(linewidth = 1.2) +
    geom_vline(xintercept = c(0.3, 0.5, 0.8), linetype = "dashed") +
    scale_x_continuous(breaks = seq(0, 1, 0.1)) +
    labs(x = "Wartość prawdopodobieństwa p", y = "Długość przedziału ufności") +
    facet_wrap(~ N) +
    ggtitle(paste("Długość przedziału ufności (n =", n_val, ")"))

  print(p_length)
}
```

Kod do wizualizacji wyników na wykresach

- Wyniki

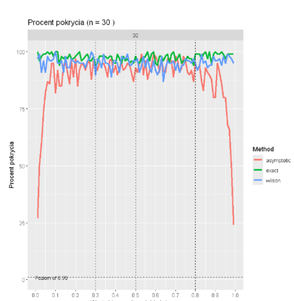
  - Porównanie wartości p dla teoretycznych i empirycznych wartości

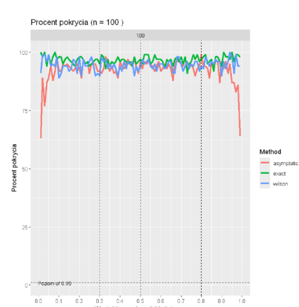
	Method	N	P	Coverage	Length
1	exact	30	0.01	100	0.13125821
2	asymptotic	30	0.01	27	0.03618798
3	wilson	30	0.01	97	0.12334693
181	exact	30	0.21	99	0.30984262
182	asymptotic	30	0.21	85	0.25285543
183	wilson	30	0.21	96	0.26162620
361	exact	30	0.41	99	0.36108257
362	asymptotic	30	0.41	90	0.31413773
363	wilson	30	0.41	95	0.31169312
721	exact	30	0.81	99	0.29966582
722	asymptotic	30	0.81	92	0.26066913
723	wilson	30	0.81	92	0.24232497

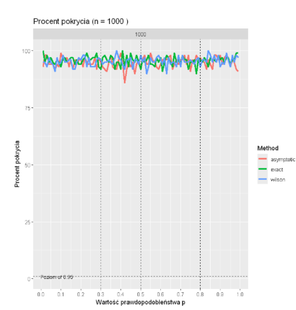
	Method	N	P	Coverage	Length
4	exact	100	0.01	100	0.05163140
5	asymptotic	100	0.01	63	0.03014480
6	wilson	100	0.01	91	0.04537877
184	exact	100	0.21	98	0.16422582
185	asymptotic	100	0.21	89	0.14114986
186	wilson	100	0.21	90	0.14007255
364	exact	100	0.41	97	0.19383334
365	asymptotic	100	0.41	94	0.18025445
366	wilson	100	0.41	92	0.17356022
724	exact	100	0.81	98	0.15608120
725	asymptotic	100	0.81	94	0.14532375
726	wilson	100	0.81	93	0.13983233

	Method	N	P	Coverage	Length
7	exact	1000	0.01	100	0.01304778
8	asymptotic	1000	0.01	93	0.01185796
9	wilson	1000	0.01	95	0.01191367
187	exact	1000	0.21	96	0.04912118
188	asymptotic	1000	0.21	93	0.04712695
189	wilson	1000	0.21	98	0.04940388
367	exact	1000	0.41	100	0.06184432
368	asymptotic	1000	0.41	95	0.05789494
369	wilson	1000	0.41	94	0.05714926
727	exact	1000	0.81	95	0.04709065
728	asymptotic	1000	0.81	95	0.04619525
729	wilson	1000	0.81	96	0.04651503

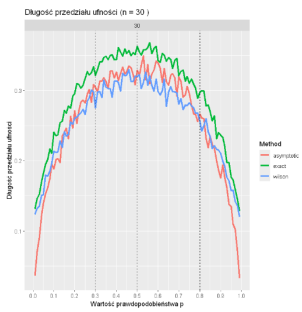
  - Porównanie procentu pokrycia jako funkcja parametru p

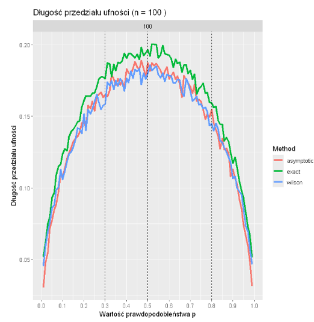


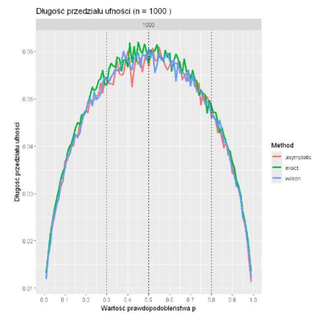




- Porównanie długości przedziału ufności jako funkcja parametru p







- Wnioski

  - Analiza procentu pokrycia przedziałów ufności:
    Analizując przedstawione wykresy prawdopodobieństwa pokrycia przedziałów ufności, możemy wyciągnąć wniosek, że wraz ze wzrostem rozmiaru próbki n, przedziały są bardziej wiarygodne i częściej pokrywają się z założonym poziomem ufności 0.95. Pomimo tego ogólnego trendu, różnice między zastosowanymi metodami są zauważalne. Metoda Walda wykazuje tendencję do gorszego przybliżania założonego poziomu ufności w porównaniu do innych metod. Metoda Cloppera-Pearsona prezentuje się najlepiej, zdecydowana większość jej wyników znajduje się powyżej poziomu 0.95.

  - Analiza długości przedziałów ufności:
    Krótsze przedziały oznaczają bardziej precyzyjne oszacowania. Z naszych wykresów wynika, że zwiększanie rozmiaru próbki n powoduje zmniejszenie długości przedziałów, wskazuje to na wzrost precyzji estymacji z większą liczbą obserwacji.

    Na naszych wykresach metoda Cloppera-Pearsona dla każdej analizowanej wartości n generuje najdłuższe przedziały. Może to wskazywać na nadmierną ostrożność tej metody, co zapewnia większe pokrycie, ale kosztem precyzji. Inne metody, które dają zbliżone długości przedziałów, mogą być bardziej odpowiednie, jeśli chcemy zrównoważyć pokrycie i precyzję.

  - Z tych obserwacji wynika, że decyzja o wyborze metody zależeć będzie od priorytetów osoby: czy ważniejsza jest precyzja estymacji (krótsze przedziały) czy pewność pokrycia prawdziwej wartości (większy procent pokrycia). Metoda Cloppera-Pearsona, choć może dawać dłuższe przedziały, prezentuje się jako metoda dająca bardziej regularne pokrycie, co może być szczególnie ważne przy skrajnych wartościach parametru p.

– Jeśli chodzi o kompromis między szerokością przedziału a częstością jego pokrycia prawdziwej wartości parametru, najodpowiedniejszą wydaje się nam metoda Wilsona.

9 Zadanie 11

- Polecenie
  Dla danych z pliku "ankieta.csv" przyjmując $\alpha = 0.05$, zweryfikuj 5 poniższych hipotez i sformułuj wnioski.

- Hipotezy

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

```
dane_kobiet <- dane$PŁEĆ[dane$PŁEĆ == "K"] # liczba kobiet wynosi 71
liczba_kobiet <- length(dane_kobiet)
print(liczba_kobiet)

library(binom)
test1 <- binom.test(liczba_kobiet , n , p=0.5, alternative = "two.sided", conf.level = 0.95)
print(test1)
```

```
	Exact binomial test

data:  liczba_kobiet and n
number of successes = 71, number of trials = 200, p-value = 4.973e-05
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2887838 0.4255862
sample estimates:
probability of success 
                 0.355 
```

P-wartość wynosi mniej niż poziom istotności, zatem możemy odrzucić tą hipotezę i przyjąć że nie jest prawdziwa (dodatkowo zauważamy, że prawdopodobienstwo tego ze hipoteza zerowa jest prawdziwa wynosi 0.355 co potwierdza odrzucenie hipotezy).

2. Prawdopodobienstwo, że pracownik jest zadowolony ze swojego wynagrodzenia w pierwszym badanym okresie jest wieksze bądź równe 0.7.

```
dane_zadowolonych <- dane$PYT_1[dane$PYT_1 >0] #ilosc zadowolonych qynosi 129
liczba_zadowolonych <- length(dane_zadowolonych)
test2 <- binom.test(liczba_zadowolonych , n , p=0.7, alternative = "greater", conf.level = 0.95)
print(test2)
```

```
	Exact binomial test

data:  liczba_zadowolonych and n
number of successes = 129, number of trials = 200, p-value = 0.9604
alternative hypothesis: true probability of success is greater than 0.7
95 percent confidence interval:
 0.5854885 1.0000000
sample estimates:
probability of success 
                 0.645 
```

P-wartość jest większa niż poziom istotności, zatem nie ma podstawy do odrzucenia hipotezy.

3. Prawdopodobienstwo, że kobieta pracuje na stanowisku menedżerskim jest równe prawdopodobienstwu, ze mężczyzna pracuje na stanowisku menedżerskim.

```
dane_menadzerskie <- dane$PŁEĆ[dane$DZIAŁ == "DK"]
n_menadzerskie <- length(dane_menadzerskie)
dane_menadzerskie_kobiety <- length(dane_menadzerskie[dane_menadzerskie == "K"])
dane_menadzerskie_faceci <- length(dane_menadzerskie[dane_menadzerskie == "M"])
x <- c(dane_menadzerskie_kobiety , dane_menadzerskie_faceci)
nn <- c(n_menadzerskie, n_menadzerskie)
test3 <- prop.test(x , nn , alternative = "t" , correct = FALSE)
test4 <- prop.test(x , nn , alternative = "t" , correct = TRUE)
print(test3) #p-value = 2.2e-16
print(test4) #p-value = 3.861e-16
```

```
	2-sample test for equality of proportions without continuity correction

data:  x out of nn
X-squared = 68.653, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.7046828 -0.4789907
sample estimates:
   prop 1    prop 2 
0.2040816 0.7959184 
```

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

data:  x out of nn
X-squared = 66.306, df = 1, p-value = 3.861e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.7148868 -0.4687866
sample estimates:
   prop 1    prop 2 
0.2040816 0.7959184 
```

P-wartość jest mniejsza niż poziom istotności, zatem odrzucamy hipotezę. Opcja zwracania nie wpływa istotnie na przyjęcie/odrzucenie hipotezy.

4. Prawdopodobienstwo, że kobieta jest zadowolona ze swojego wynagrodzenia w pierwszym badanym okresie jest równe prawdopodobienstwu, że mężczyzna jest zadowolony ze swojego wynagrodzenia w pierwszym badanym okresie.

```
dane_kobiety <- dane$PYT_1[dane$PŁEĆ == "K"]
dane_faceci <- dane$PYT_1[dane$PŁEĆ == "M"]
liczba_kobiety <- length(dane_kobiety)
liczba_facetow <- length(dane_faceci)
dane_kobiet_zadowolenie <-  length(dane_kobiety[dane_kobiety > 0])
dane_facet_zadowolenie <-  length(dane_faceci[dane_faceci > 0])
xx <- c(dane_kobiet_zadowolenie , dane_facet_zadowolenie)
nnn <- c(liczba_kobiety , liczba_facetow)

test5 <- prop.test(xx , nnn , alternative = "t" , correct = FALSE)
test6 <- prop.test(xx , nnn , alternative = "t" , correct = TRUE)
print(test5) #p-value = 0.7098
print(test6) #p-value = 0.8277
```

```
	2-sample test for equality of proportions without continuity correction

data:  xx out of nnn
X-squared = 0.13847, df = 1, p-value = 0.7098
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1115402  0.1641660
sample estimates:
   prop 1    prop 2
0.6619718 0.6356589
```

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

data:  xx out of nnn
X-squared = 0.047399, df = 1, p-value = 0.8277
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1224584  0.1750843
sample estimates:
   prop 1    prop 2
0.6619718 0.6356589
```

P-wartość w obu testach jest większa niż poziom istotności, zatem nie ma podstawy do odrzucenia hipotezy (ponownie opcja zwracania nie wpływa istotnie na przyjęcie/odrzucenie hipotezy).

5. Prawdopodobienstwo, że kobieta pracuje w dziale obsługi kadrowo-płacowej jest większe lub równe prawdopodobienstwu, że ˙ mężczyzna pracuje w dziale obsługi kadrowo- płacowej.

```
dane_hr <- dane$PŁEĆ[dane$DZIAŁ == "HR"]
liczba_kobiety_hr <- length(dane_hr[dane_hr == "K"])
liczba_faceci_hr <- length(dane_hr[dane_hr == "M"])
ilosc_hr <- length(dane_hr)
xxx <- c(liczba_kobiety_hr , liczba_faceci_hr)
nnnn <- c(ilosc_hr , ilosc_hr)

test7 <- prop.test(xxx , nnnn , alternative = "less" , correct = FALSE)
test8 <- prop.test(xxx , nnnn , alternative = "less" , correct = TRUE)
print(test7) #p-value = 2.579e-09
print(test8) #p-value = 1.148e-08
```

```
        2-sample test for equality of proportions without continuity
        correction

data:  xxx out of nnnn
X-squared = 34.129, df = 1, p-value = 2.579e-09
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000 -0.6018763
sample estimates:
   prop 1    prop 2
0.1290323 0.8709677
```

```
data:  xxx out of nnnn
X-squared = 31.226, df = 1, p-value = 1.148e-08
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000 -0.5696182
sample estimates:
   prop 1    prop 2
0.1290323 0.8709677
```

P-wartość w obu testach jest mniejsza niż poziom istotności, zatem odrzucamy hipotezę (znów opcja zwracania nie wpływa istotnie na przyjęcie/odrzucenie hipotezy).

10 Zadanie 12

- Polecenie
  Wyznacz symulacyjnie moc testu dokładnego oraz moc testu asymptotycznego w przypadku weryfikacji hipotezy zerowej $H_0 : p = 0.9$ przeciwko $H1 : p \neq 0.9$ przyjmując wartość $1\alpha = 0.95$. Uwzględnij różne wartości alternatyw i różne rozmiary próby. Sformułuj wnioski.

- Niestety z powodu problemów z wygenerowaniem wszystkich wykresów w pętli dla wszystkich n z powodu zbyt dużego obciążenia komputera, byłyśmy zmuszone nasz kod wywoływać osobno dla każdej z wartości n. Wklejamy tylko kod dla $n = 1000$, ale oba pozostałe wyglądały tak samo.

- Opis kodu
  Tworzymy wektor wartosci_p z wartościami od 0.01 do 0.99, wyłączając wartość 0.9 z krokiem co 0.01. Następnie ustalamy liczbę powtórzeń N na 500. Tworzymy wektory n_power_binom, n_power_asymptotyczny_z_poprawka
  i n_power_asymptotyczny_bez_poprawki do przechowywania wyników mocy testu dla funkcji binom.test, prop.test oraz z funckji prop.test z poprawką na ciągłość. Później generujemy dane z rozkładu dwumianowego dla próby o wielkości n z prawdopodobieństwem sukcesu p. Następnie przeprowadzamy testy, zliczamy ile razy hipoteza zerowa została odrzucona w każdym z testów. Na drugim zdjęciu zapisujemy wyniki w ramce danych. Na trzecim zdjęciu wizualizujemy wyniki

na 3 wykresach (osobno dla każdego n). Każdy test ma inny kolor na wykresie: czerwony - test binom, niebieski - test asymptotyczny z poprawką, fioletowy - test asymptotyczny bez poprawki.

- Kody

```r
N <- 500
alpha <- 0.05
wartosci_p <- c(seq(0.01, 0.89, by = 0.01), seq(0.91, 0.99, by = 0.01))
wyniki <- list()
n <- 1000
n_power_binom <- vector("numeric", length(wartosci_p))
n_power_aymptotyczny_z_poprawka <- vector("numeric", length(wartosci_p))
n_power_asymptotyczny_bez_poprawki <- vector("numeric", length(wartosci_p))
for (i in 1:length(wartosci_p)) {
  p_konkretne <- wartosci_p[i]
  odrzuc_binom <- 0
  odrzuc_asymptotyczny_z_poprawka <- 0
  odrzuc_asymptotyczny_bez_poprawki <- 0
  for (k in 1:N) {
    data <- rbinom(n, 1, 0.5)
    binom_test <- binom.test(sum(data), n, p = p_konkretne, alternative = "t")
    prop_test_z <- prop.test(sum(data), n, p = p_konkretne, alternative = "t", correct = TRUE)
    prop_test_bez <- prop.test(sum(data), n, p = p_konkretne, alternative = "t", correct = FALSE)
    if (binom_test$p.value < alpha) {
      odrzuc_binom <- odrzuc_binom + 1
    }
    if (prop_test_z$p.value < alpha) {
      odrzuc_asymptotyczny_z_poprawka <- odrzuc_asymptotyczny_z_poprawka + 1
    }
    if (prop_test_bez$p.value < alpha) {
      odrzuc_asymptotyczny_bez_poprawki <- odrzuc_asymptotyczny_bez_poprawki + 1
    }
  }
  n_power_binom[i] <- odrzuc_binom / N
  n_power_aymptotyczny_z_poprawka[i] <- odrzuc_asymptotyczny_z_poprawka / N
  n_power_asymptotyczny_bez_poprawki[i] <- odrzuc_asymptotyczny_bez_poprawki / N
}
```

Główna pętla kodu

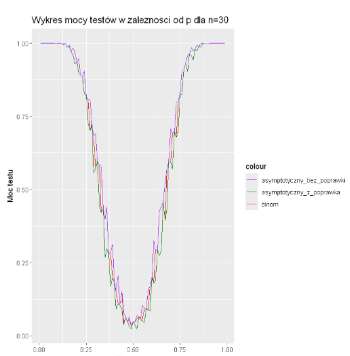
```r
df <- data.frame(p_konkretne = wartosci_p,
                 power_binom = n_power_binom,
                 power_asymptotyczny_z_poprawka = n_power_aymptotyczny_z_poprawka,
                 power_asymptotyczny_bez_poprawki = n_power_asymptotyczny_bez_poprawki,
                 n = as.factor(n))
wyniki[[as.character(n)]] <- df
```

Zapisywanie wyników do ramki danych

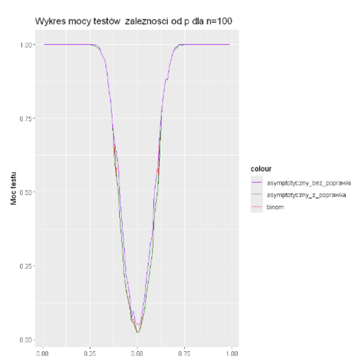
```r
library(ggplot2)

plot <- ggplot(data = do.call(rbind, wyniki),
               aes(x = p_konkretne)) +
  geom_line(aes(y = power_binom, color = "binom")) +
  geom_line(aes(y = power_asymptotyczny_z_poprawka, color = "asymptotyczny_z_poprawka")) +
  geom_line(aes(y = power_asymptotyczny_bez_poprawki, color = "asymptotyczny_bez_poprawki")) +
  labs(x = "Wartosc p", y = "Moc testu") +
  scale_color_manual(values = c("binom" = "red", "asymptotyczny_z_poprawka" = "#0b780b", "asymptotyczny_bez_poprawki" = "purple")) +
  ggtitle("Wykres mocy testów w zaleznosci od p dla n=1000")

print(plot)
```

Wygenerowanie wykresów wizualizujących wynik

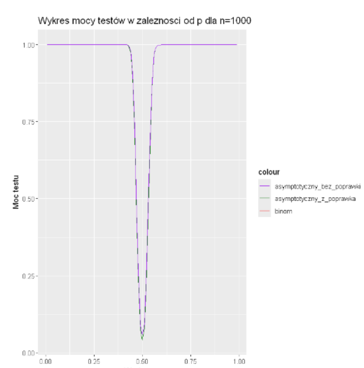
- Wyniki



n=30



n=100



n=1000