

SPRAWOZDANIE 1

Analiza danych ankietowych

Aleksandra Grzeszczuk
album 255707

Jacek Wszola
album 255718

7 listopada 2021

Spis treści

1	Lista 1	1
1.1	Zadanie 1	1
1.1.1	Podpunkt A	1
1.1.2	Podpunkt B	3
1.1.3	Podpunkt C	3
1.1.4	Podpunkt D	4
1.1.5	Podpunkt E	5
1.1.6	Podpunkt F	5
1.2	Zadanie 2	6
2	Lista 2 i 3	8
2.1	Zadanie 1	8
2.2	Zadanie 2	8
2.3	Zadanie 3	11
3	Lista 4 i 5	14
3.1	Zadanie 1	14
3.2	Zadanie 2	14
3.3	Zadanie 3	16

1 Lista 1

1.1 Zadanie 1

Po wczytaniu danych z pliku `Choroba.csv` zmodyfikujemy nasze dane, tak abyśmy mogli operować na nich w łatwiejszy sposób. Łatwo sprawdzić, że mamy do czynienia z danymi typu `integer`. O wiele prościej będzie posługiwać się zmiennymi typu `factor`. Zmienne ilościowe – w przeciwieństwie do zmiennych jakościowych – są mierzalne i możliwe do uporządkowania.

```
dane <- mutate(dane, across(c("STATUS", "SEKTOR", "CHORY_ZD", "OSZCZED"),
                             as.factor))
```

Zmienimy jeszcze nazwy zmiennych, aby prościej tworzyć legendy przy wykresach.

```
dane <- dane %>% mutate(STATUS = fct_recode(STATUS, "Wysoki" = "1",
                                             "Średni" = "2", "Niski" = "3"))
dane <- dane %>% mutate(OSZCZED = fct_recode(OSZCZED,
                                             "Posiada oszczędności" = "1",
                                             "Nie posiada oszczędności" = "0"))
dane <- dane %>% mutate(CHORY_ZD = fct_recode(CHORY_ZD, "Chory" = "1",
                                             "Zdrowy" = "0"))
```

1.1.1 Podpunkt A

Naszym zadaniem jest sporządzenie tablic liczości dla zmiennych `Oszczędności` oraz `Chory/Zdrowy`, biorąc pod uwagę wszystkie dane, jak również w podgrupach: ze względu na zmienne `Status`, `Sektor` (razem i z osobna).

	SEKTOR	ZDROWY	CHORY
1	I	95	22
2	II	45	34

Tabela 1: Zestawienie zmiennych CHORY/ZDROWY i SEKTOR

Widzimy, że w sektorze pierwszym mieszka 117 osób, z czego 81.2% jest zdrowych. Sektor drugi zamieszkuje jedynie 79 osób i tylko 57% z nich jest zdrowych. Można na tej podstawie wywnioskować, że osoby z sektora II mają utrudniony dostęp do lekarzy, przychodni, mieszkają z dala od szpitali.

	STATUS	ZDROWY	CHORY
1	Wysoki	53	24
2	Średni	36	13
3	Niski	51	19

Tabela 2: Zestawienie zmiennych CHORY/ZDROWY i STATUS

W badanej grupie najwięcej jest osób o wysokim statusie – aż 77 z czego jedynie 68% zdrowych. Osób ze średnim statusem jest najmniej, bo jedynie 49, ale to właśnie ta grupa ma

największy odsetek osób zdrowych równy 73.5%. Grupa osób ze statusem niskim (70) posiada również wysoki odsetek osób zdrowych, bo 72.9%. Możemy stąd wywnioskować, że osoby o statusie niskim bądź średnim mają mniej zmartwień, przez co są zdrowsze od osób o wysokim statusie – zawzyczają pracoholików.

	STATUS	OSZCZĘDNOŚCI	BRAK OSZCZĘDNOŚCI
1	Wysoki	60	17
2	Średni	24	25
3	Niski	22	48

Tabela 3: Zestawienie zmiennych STATUS i OSZCZĘDNOŚCI

Nie trzeba analizować wartości procentowych, żeby zauważyć, że prawie wszystkie osoby o wysokim statusie posiadają oszczędności. W pozostałych grupach zdecydowana większość ich nie gromadzi. Możemy stąd wywnioskować, że nie mają z czego odkładać bądź wydają zbyt pochopnie, nie patrząc na ewentualne przyszłe wydatki.

	SEKTOR	MA OSZCZED	NIE MA OSZCZED
1	I	52	65
2	II	54	25

Tabela 4: Tabela wielodzielcza uwzględniająca zmienną SEKTOR oraz OSZCZĘDNOŚCI

Jedynie 44.4% osób z pierwszego sektora posiada oszczędności, podczas gdy z drugiego jest to aż 68.4%. Jest to dość zaskakujące, ponieważ, jak już wyżej przeanalizowaliśmy, zdecydowana większość osób zdrowych mieszka w sektorze I.

	STATUS	SEKTOR	ZDROWY	CHORY
1	Wysoki	1	31	7
2	Wysoki	2	22	17
3	Średni	1	23	3
4	Średni	2	13	10
5	Niski	1	41	12
6	Niski	2	10	7

Tabela 5: Zestawienie zmiennych STATUS, SEKTOR i CHORY/ZDROWY

	STATUS	SEKTOR	OSZCZĘDNOŚCI	BRAK OSZCZĘDNOŚCI
1	Wysoki	1	27	11
2	Wysoki	2	33	6
3	Średni	1	11	15
4	Średni	2	13	10
5	Niski	1	14	39
6	Niski	2	8	9

Tabela 6: Zestawienie zmiennych STATUS, SEKTOR i OSZCZĘDNOŚCI

1.1.2 Podpunkt B

Sporządzamy tabelę wielodzietczą uwzględniającą zmienne Chory/Zdrowy oraz Sektor.

	SEKTOR	ZDROWY	CHORY
1	I	95	22
2	II	45	34

```
dane%>%group_by(SEKTOR)%>%frq(CHORY_ZD)

##
## CHORY_ZD <categorical>
## # grouped by: 1
## # total N=117  valid N=117  mean=1.19  sd=0.39
##
## Value | N | Raw % | Valid % | Cum. %
## -----
## Zdrowy | 95 | 81.20 | 81.20 | 81.20
## Chory  | 22 | 18.80 | 18.80 | 100.00
## <NA>   | 0  | 0.00  | <NA>  | <NA>
##
##
## CHORY_ZD <categorical>
## # grouped by: 2
## # total N=79  valid N=79  mean=1.43  sd=0.50
##
## Value | N | Raw % | Valid % | Cum. %
## -----
## Zdrowy | 45 | 56.96 | 56.96 | 56.96
## Chory  | 34 | 43.04 | 43.04 | 100.00
## <NA>   | 0  | 0.00  | <NA>  | <NA>
```

1.1.3 Podpunkt C

Ponownie sporządzamy tabelę wielodzietczą, uwzględniającą tym razem zmienne Chory/Zdrowy oraz Status.

	STATUS	ZDROWY	CHORY
1	Wysoki	53	24
2	Średni	36	13
3	Niski	51	19

```
dane%>%group_by(STATUS)%>%frq(CHORY_ZD)

##
## CHORY_ZD <categorical>
```

```
## # grouped by: Wysoki
## # total N=77 valid N=77 mean=1.31 sd=0.47
##
## Value | N | Raw % | Valid % | Cum. %
## -----
## Zdrowy | 53 | 68.83 | 68.83 | 68.83
## Chory | 24 | 31.17 | 31.17 | 100.00
## <NA> | 0 | 0.00 | <NA> | <NA>
##
##
## CHORY_ZD <categorical>
## # grouped by: Średni
## # total N=49 valid N=49 mean=1.27 sd=0.45
##
## Value | N | Raw % | Valid % | Cum. %
## -----
## Zdrowy | 36 | 73.47 | 73.47 | 73.47
## Chory | 13 | 26.53 | 26.53 | 100.00
## <NA> | 0 | 0.00 | <NA> | <NA>
##
##
## CHORY_ZD <categorical>
## # grouped by: Niski
## # total N=70 valid N=70 mean=1.27 sd=0.45
##
## Value | N | Raw % | Valid % | Cum. %
## -----
## Zdrowy | 51 | 72.86 | 72.86 | 72.86
## Chory | 19 | 27.14 | 27.14 | 100.00
## <NA> | 0 | 0.00 | <NA> | <NA>
```

1.1.4 Podpunkt D

Naszym kolejnym zadaniem jest przeprowadzenie kategoryzacji zmiennej *Wiek*. Kategoryzacja zmiennych nazywana jest również grupowaniem – jest to proces budowania nowych zmiennych na podstawie już istniejących.

```
dane <- mutate(dane, KATEGORIA_WIEKOWA =
  cut(WIEK, breaks = c(0, 13, 28, 42, 62, Inf)))
```

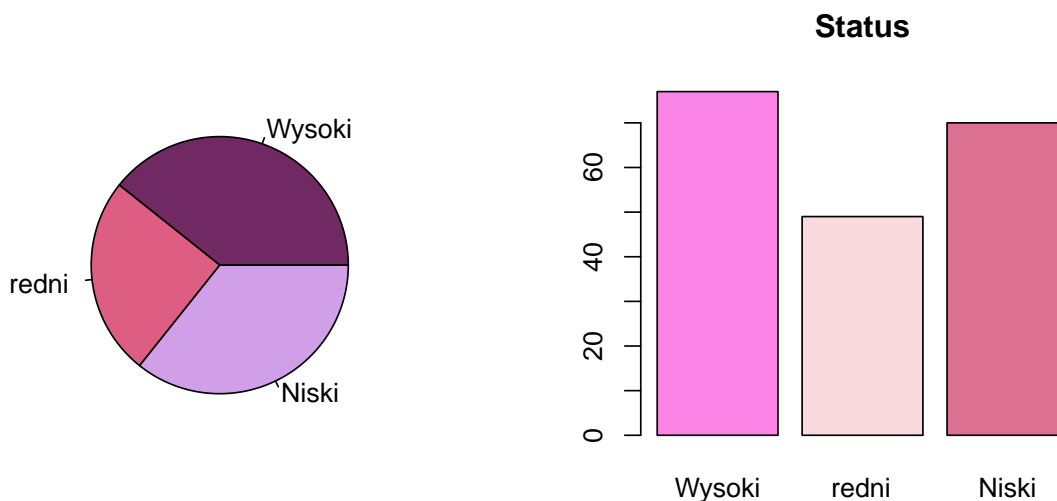
KATEGORIA WIEKOWA	<i>n</i>
≤ 13	64
13-28	62
28-42	38
42-62	19
≥ 62	13

Tabela 7: Kategoryzacja zmiennej wiek

Analizując powyższą tabelę, widzimy, że najwięcej jest osób w wieku do 28 roku życia. Powyżej, liczba osób stopniowo maleje. Średnia wieku w Polsce wynosi około 75 lat, zatem nie dziwi nas, że osób powyżej 62 roku życia jest najmniej – jedynie 13.

1.1.5 Podpunkt E

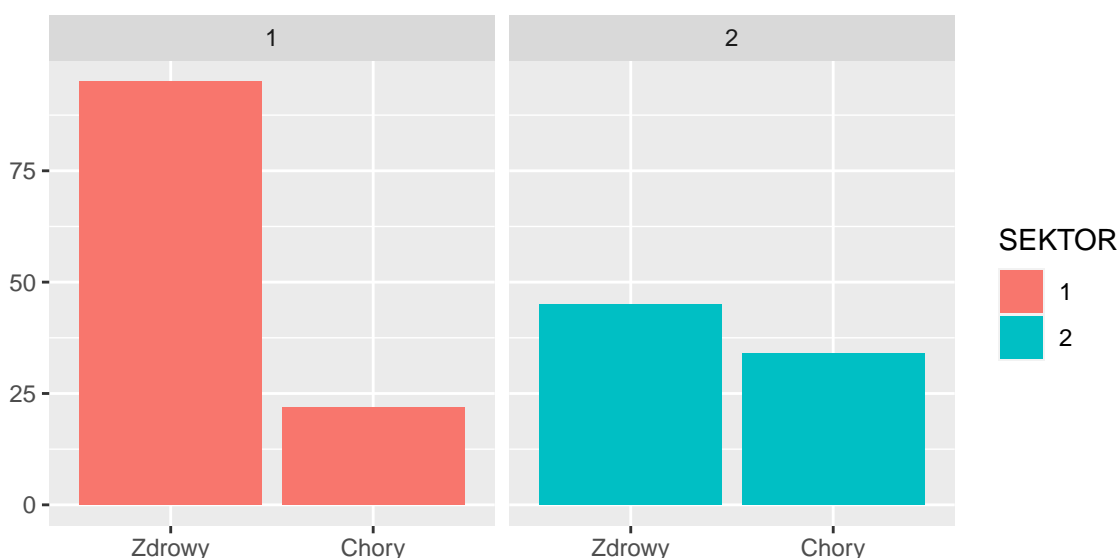
Skupimy się teraz na zmiennej **Status**. Poniżej prezentujemy jej wykresy; kołowy i słupkowy.



Spoglądając na wykres kołowy, możemy domyślać się, że osób o statusie wysokim oraz niskim jest tyle samo, zaś ze średnim – niewiele mniej. Diagramy kołowe nie są jednak zbyt dokładne, dlatego w precyzyjniejszej analizie pomagają nam wykres słupkowy, który ostatecznie informuje, że osób z wysokim oraz niskim statusem jest najwięcej zaś ze średnim zdecydowanie mniej.

1.1.6 Podpunkt F

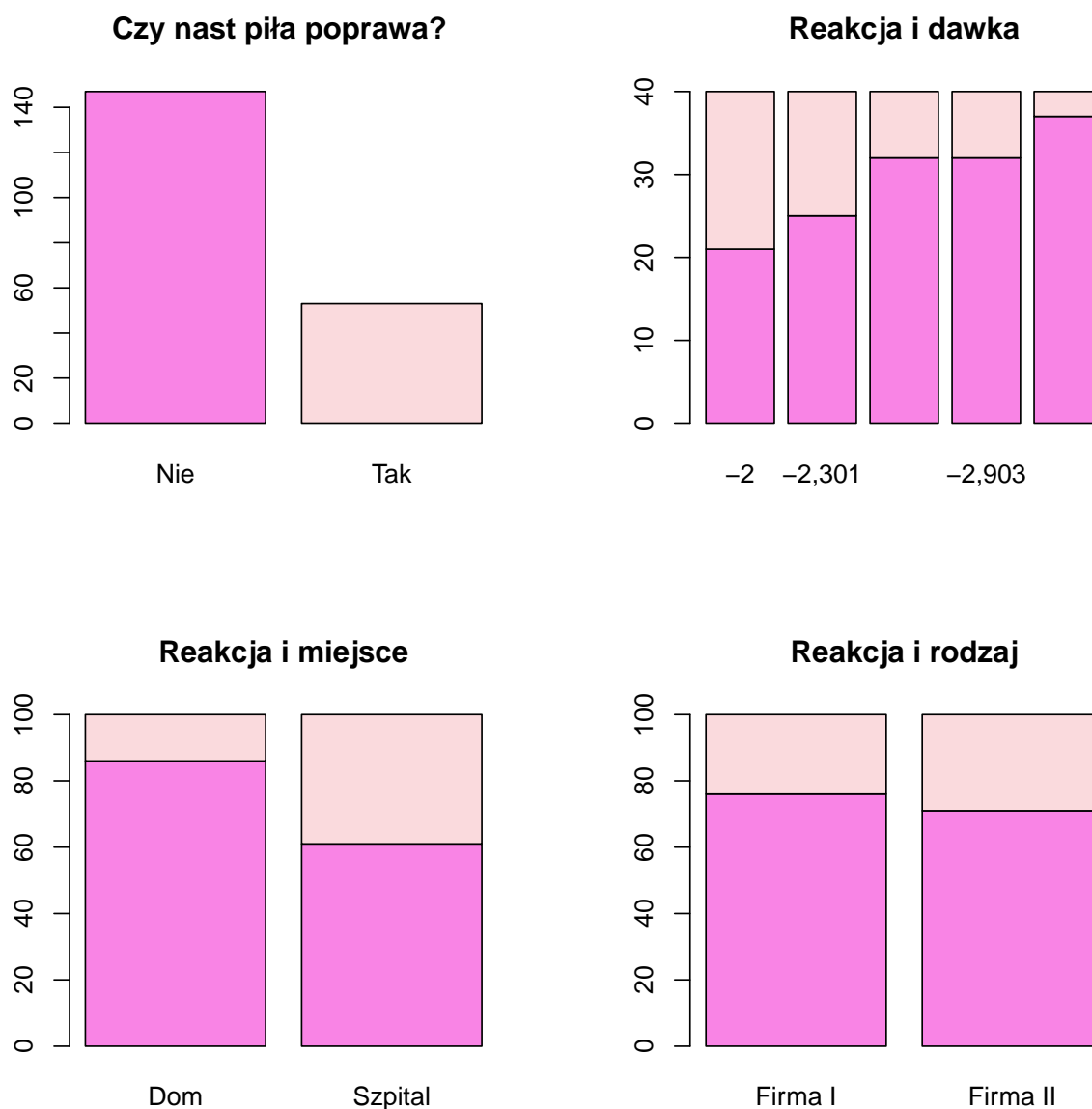
Sporządzimy teraz skategoryzowane wykresy zmiennej **Chory/Zdrowy** przyjmując za zmienną kategoryzującą zmienną **Sektor**.



Widzimy, że w sektorze I mieszka zdecydowanie więcej osób zdrowych oraz zdecydowanie mniej chorych niż w sektorze II.

1.2 Zadanie 2

Dane w pliku `Reakcja.csv` zawierają informacje o reakcji na lek. Naszym zadaniem będzie sporządzenie wykresów zmiennej `Reakcja` w całej badanej grupie i w podgrupach ze względu na kategorie pozostałych zmiennych. Poniżej prezentujemy odpowiednie wykresy.



```
zad2%>%group_by(Miejsce)%>%frq(Reakcja)

##
## Reakcja <category>
## # grouped by: Dom
## # total N=100 valid N=100 mean=1.14 sd=0.35
```

```
##
## Value | N | Raw % | Valid % | Cum. %
## -----
## Nie | 86 | 86.00 | 86.00 | 86
## Tak | 14 | 14.00 | 14.00 | 100
## <NA> | 0 | 0.00 | <NA> | <NA>
##
##
## Reakcja <ategorical>
## # grouped by: Szpital
## # total N=100 valid N=100 mean=1.39 sd=0.49
##
## Value | N | Raw % | Valid % | Cum. %
## -----
## Nie | 61 | 61 | 61 | 61
## Tak | 39 | 39 | 39 | 100
## <NA> | 0 | 0 | <NA> | <NA>
```

Przyglądając się powyższym wykresom oraz tabeli wielodzzielczej, widzimy, że poprawa wystąpiła w znacznej mniejszości badanej grupy. Wpływ na wystąpienie poprawy z dużym prawdopodobieństwem miało miejsce leczenia – aż 86% pacjentów leczonych w domu nie wyzdrowiało, podczas gdy w szpitalu było to „jedynie” 61%. Możemy po tym wnioskować, że w domu niekoniecznie poprawnie reagowano na różne stany chorobowe, powikłania, nie zwalczano wystarczająco objawów.

	RODZAJ	MIEJSCE	NASTĄPIŁA POPRAWA	BRAK POPRAWY
1	Firma I	Dom	5	45
2	Firma I	Szpital	19	31
3	Firma II	Dom	9	41
4	Firma II	Szpital	20	30

Tabela 8: Zależność wystąpienia poprawy

Poprawa stanu zdrowia nie zależała także od różnych rodzajów leków – wyniki z obydwu firm są prawie identyczne.

2 Lista 2 i 3

2.1 Zadanie 1

Rozważamy pewną hipotetyczną bazę danych. Za pomocą funkcji `sample()` w pakiecie `stats` wylosujemy próbkę rozmiaru około 1/10 liczby obserwacji. Rozważymy losowanie ze zwracaniem oraz bez zwracania. Obie metody przetestujemy dla konkretnych danych, w tym przypadku `survey` z biblioteki `MASS`.

```
losowanie <- function(x, zwr){
  if(zwr == TRUE){
    return(sample(x, round(length(x)/10), replace = TRUE))
  }
  else{
    return(sample(x, round(length(x)/10), replace = FALSE))
  }
}

losowanie(survey$Age, zwr = TRUE)

## [1] 17.167 20.000 19.417 18.417 18.000 19.333 16.917 23.583 19.667 17.750
## [11] 18.417 18.333 18.167 18.333 22.833 17.417 17.083 18.083 17.417 18.500
## [21] 17.583 19.917 19.833 17.917

losowanie(survey$Age, zwr = FALSE)

## [1] 18.917 19.917 17.833 18.500 17.417 18.417 17.250 19.917 18.417 27.333
## [11] 18.083 19.083 18.583 17.167 19.000 24.667 16.750 21.167 17.500 17.417
## [21] 21.250 18.500 17.750 17.167
```

2.2 Zadanie 2

Zajmiemy się teraz estymacją przedziałową parametru p z rozkładu dwumianowego $\mathcal{B}(n, p)$. W tym celu przeprowadzimy szereg symulacji i znając rzeczywistą wartość parametru p , zbędziemy dokładność różnych metod estymacji. Szczególną uwagę zwrócimy na cechy takie jak:

- prawdopodobieństwo pokrycia,
- długość przedziału ufności,
- podatność na zmianę parametrów rozkładu.

Będziemy badać trzy metody konstrukcji przedziałów ufności: Cloppera-Pearsona, Walda oraz Wilsona. We wszystkich przypadkach uwzględnimy poziom ufności równy 0.95, a w symulacjach posłużymy się funkcją `binom.confint()` z pakietu `binom`.

Przy ustalonych wartościach parametrów n, p będziemy tworzyć 10.000 prób z rozkładu $\mathcal{B}(n, p)$. Każdej próbie przyporządkujemy przedział ufności obliczony zgodnie z odpowiednią metodą. Na podstawie realizacji obliczymy średnią długość przedziałów ufności oraz prawdopodobieństwa pokrycia, zgodnie z poniższym kodem.

```

ps <- seq(0.05, 0.95, 0.05)
#n = liczba prób, m = metoda estymacji
int.test <- function(n, m){
  ps <- seq(0.05, 0.95, 0.05)
  pr.pokrycia <- c()
  sr.dlugosc <- c()
  for(p in ps){
    X <- rbinom(10000, n, p)
    intervals <- binom.confint(X, n, conf.level = 0.95, methods = m)
    lefts <- intervals$lower
    rights <- intervals$upper
    Y <- 0
    dl <- c()
    for(i in 1:10000){
      dl[i] <- rights[i] - lefts[i]
      if((lefts[i] <= p) & (p <= rights[i])){
        Y <- Y+1
      }
    }
    pr.pokrycia <- append(pr.pokrycia, Y/10000)
    sr.dlugosc <- append(sr.dlugosc, mean(dl))
  }
  return(c(pr.pokrycia, sr.dlugosc))
}

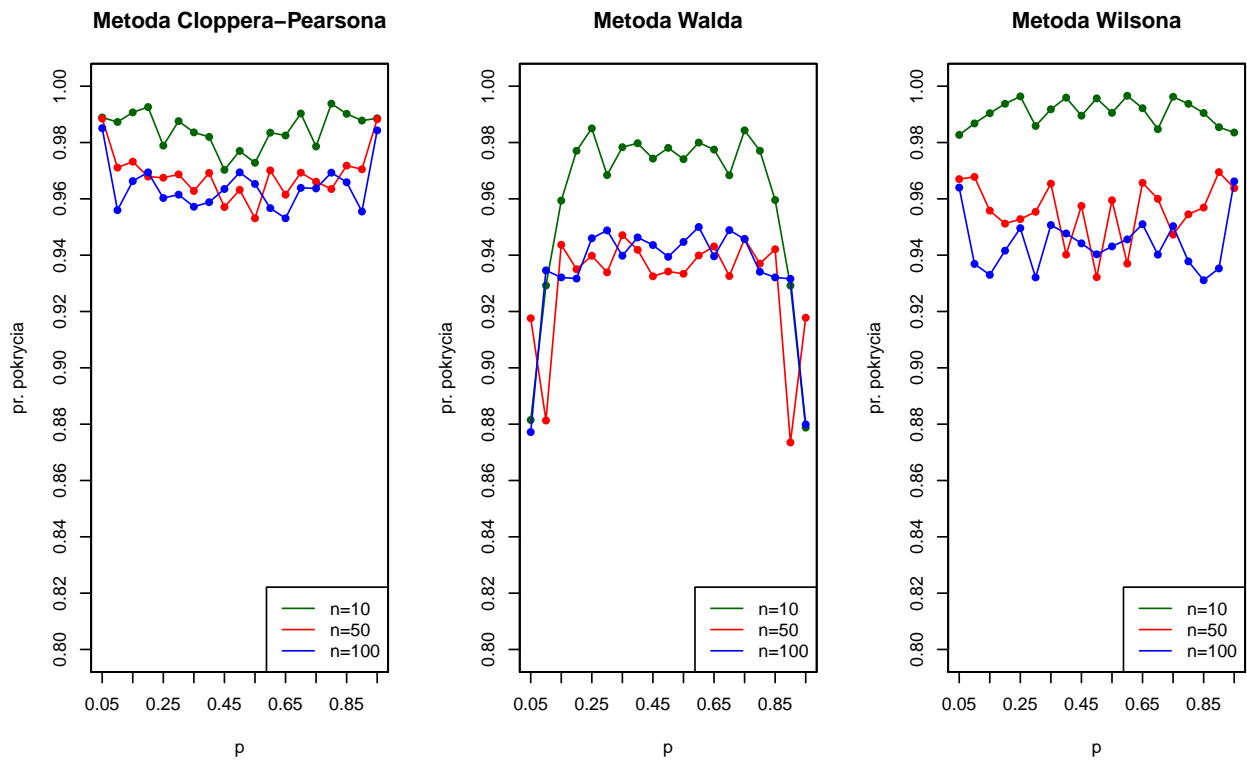
```

Spośród trzech badanych metod to metoda Cloppera-Pearsona cechuje się najwyższym prawdopodobieństwem pokrycia. Jest ono bowiem najmniej podatne na zmianę parametrów rozkładu. Największe prawdopodobieństwo pokrycia otrzymujemy przy małych wartościach n , jednak przy dużych wartościach w najgorszym wypadku nie przekracza ono 95%. Można więc powiedzieć, że dokładność tej metody jest satysfakcjonująca.

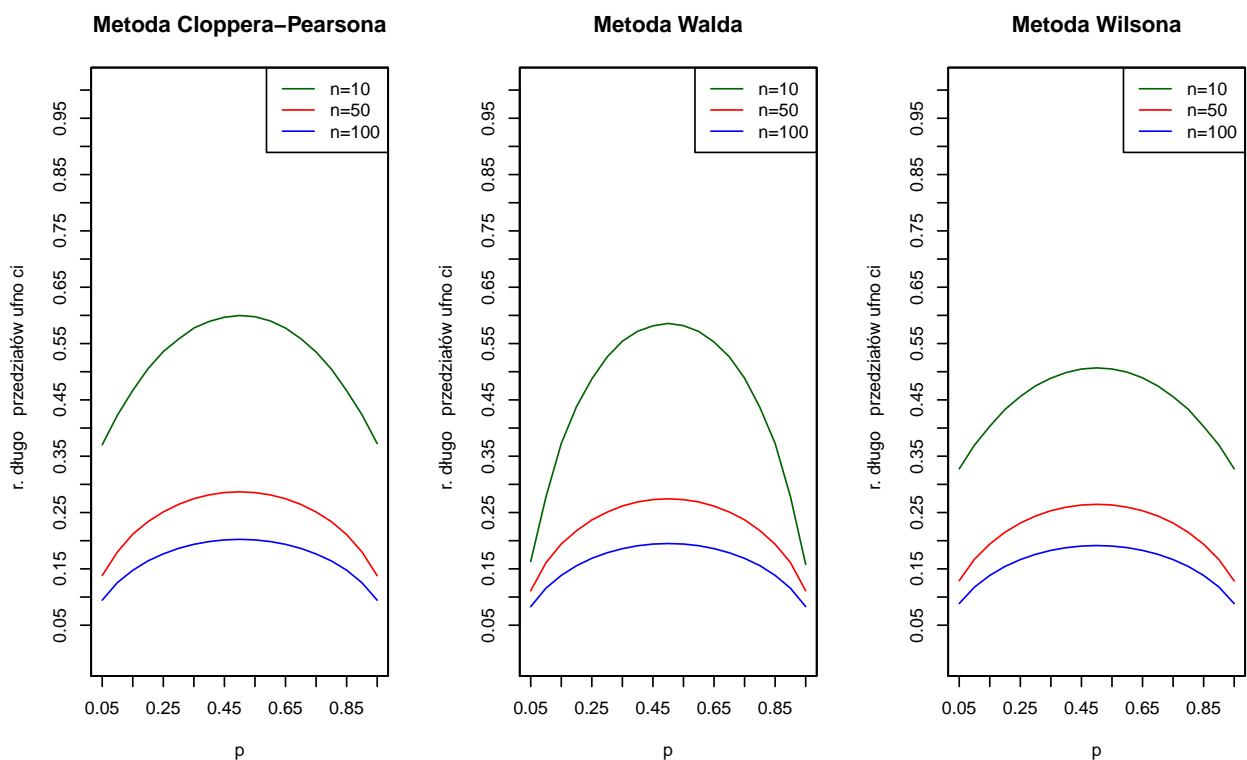
W przypadku metody Walda obserwujemy, że najgorzej radzi sobie ona dla skrajnych (bliższych 0 lub 1) wartości p . Wówczas prawdopodobieństwo pokrycia spada nawet do 87%. Warto również podkreślić, że dla $n = 50$ oraz $n = 100$ prawdopodobieństwa pokrycia są zbliżone. Podobnie jest w przypadku metody Wilsona, choć tutaj prawdopodobieństwo pokrycia nie spada poniżej 93%, a dla małych wartości n jest ono bardzo bliskie 100%. Metoda ta dla tego rzędu parametru n okazuje się efektywniejsza niż Metoda Cloppera-Pearsona, która lepiej sprawdza się większych wartości n .

Przechodząc do średniej długości przedziałów ufności, obserwujemy, że najlepiej wypada metoda Wilsona. Długość najdłuższego przedziału to około 0.5. We wszystkich przypadkach najdłuższe przedziały otrzymujemy dla małych wartości n oraz $p = 1/2$, przy zachowaniu symetrii względem parametru p . Jeśli chodzi zaś o próby większego rozmiaru, wszystkie badane metody dają podobne wyniki w granicach od 0.07 do 0.3. Metodę Walda cechują krótkie przedziały dla skrajnych wartości p , niezależnie od wartości n .

Trudno jednoznacznie określić, która z badanych metod jest najlepsza. Metoda Cloppera-Pearsona daje wysokie prawdopodobieństwa pokrycia, jednak przedziały są najdłuższe. Metoda Wilsona generuje z kolei najkrótsze przedziały, jednak dla dużych wartości n prawdopodobieństwa pokrycia mogą być niesatysfakcjonujące.



Rysunek 1: Wyniki symulacji – prawdopodobieństwa pokrycia

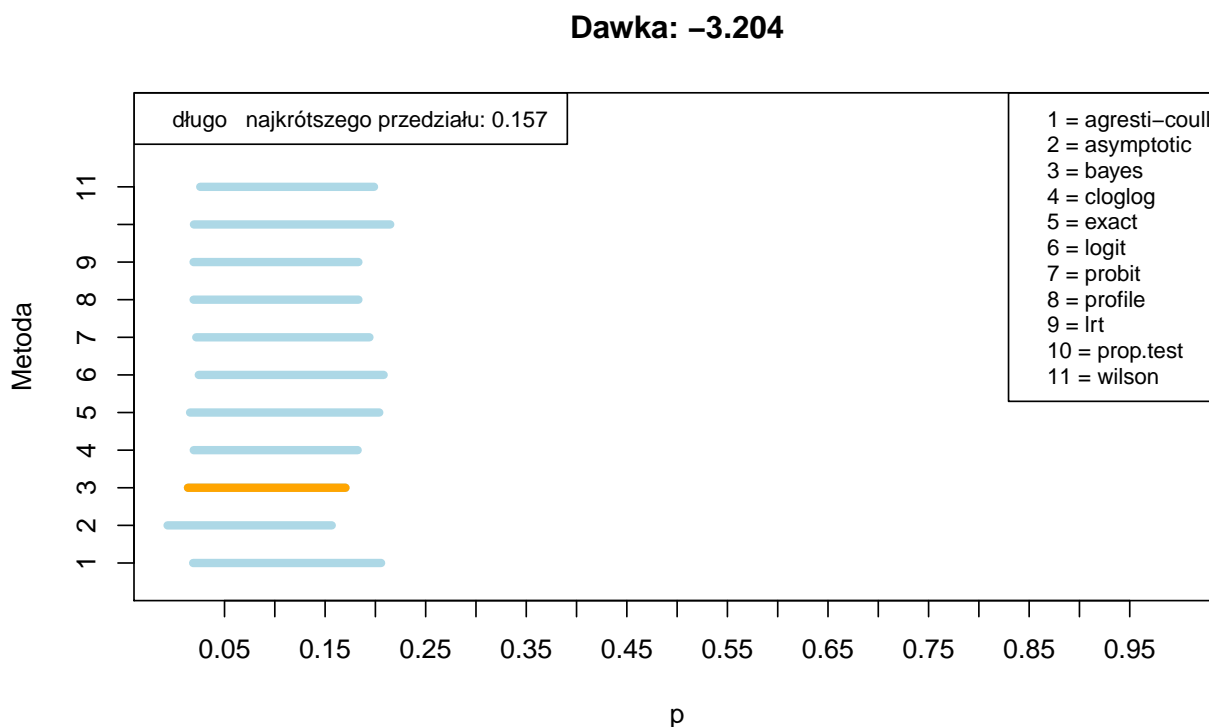


Rysunek 2: Wyniki symulacji – długość przedziałów

2.3 Zadanie 3

Informacje zdobyte z poprzedniego zadania wykorzystamy teraz do ponownego przeanalizowania pliku `reakcja.csv`, tym razem pod kątem przedziałów ufności. W podgrupach ze względu na wielkość dawki będziemy konstruować przedziały ufności, ponownie, na poziomie ufności 0.95. Porównamy wszystkie metody, które oferuje funkcja `binom.confint()`, i postaramy się wybrać najlepszą z nich.

Wyróżniliśmy w sumie 5 podgrup (ze względu na wielkość dawki) i porównaliśmy 11 metod przedziałów ufności. Wyniki zamieszczamy na poniższych wykresach.

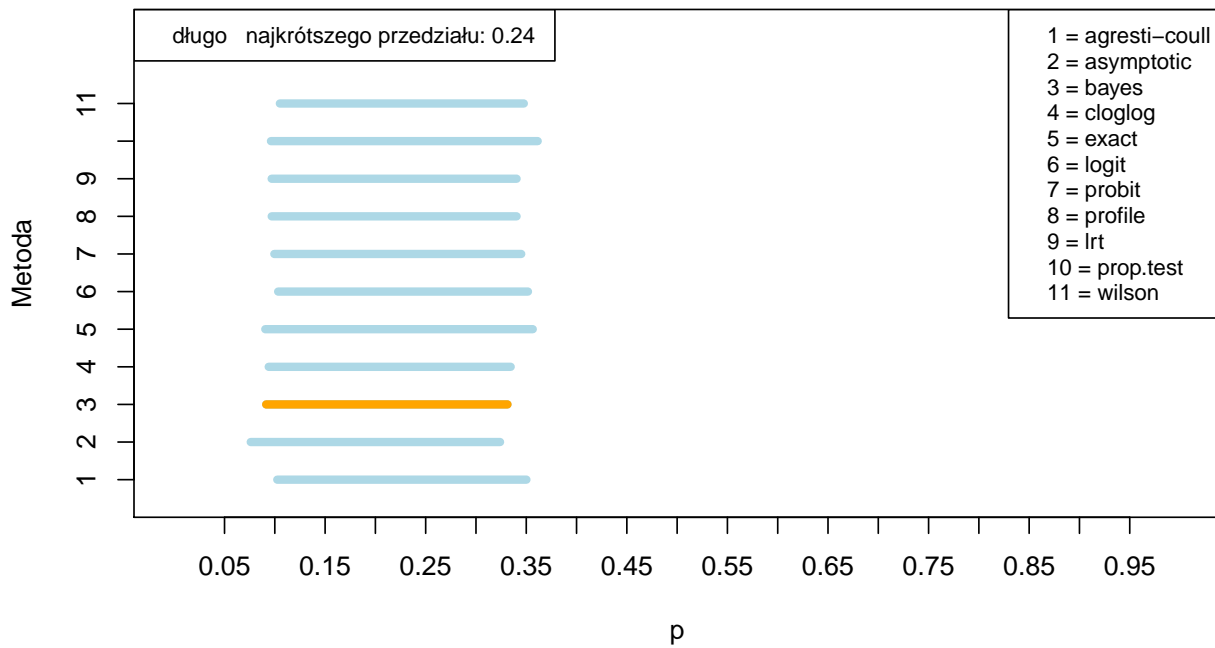


Rysunek 3: Porównanie wszystkich metod

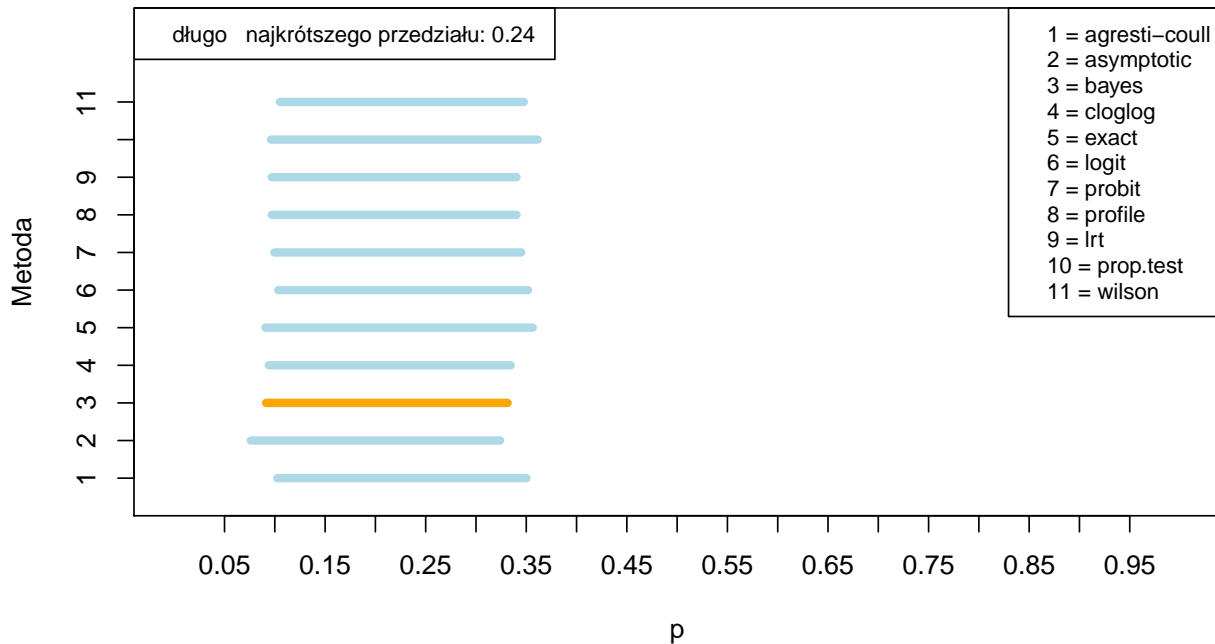
Łatwo stwierdzić, że pod względem długości przedziałów ufności na prowadzenie wysuwają się dwie metody: Bayesa (dla wyższych dawek) oraz Wilsona (dla niższych dawek). Ta obserwacja nasze ustalenia z zadania 2, gdzie stwierdziliśmy, że właśnie metoda Wilsona spośród trzech badanych daje w rezultacie najkrótsze przedziały ufności. Przyjrzyjmy się w szczególności metodom badanim w zadaniu 2 i na podstawie otrzymanych wyników spróbujmy stwierdzić, która z nich sprawdzi się najlepiej w kwestii najwyższego prawdopodobieństwa pokrycia.

Przypomnijmy, że przy klasyfikacji grup, wzięliśmy $n = 40$, zatem na wykresach z zadania 2 najbardziej interesują nas czerwone przypadki. W przypadku małych wartości p (a z takimi mamy do czynienia przy dawkach -3.204 oraz -2.903) na pewno nie wybralibyśmy metody Walda, która nie radziła sobie ze skrajnymi wartościami parametrów. Być może warto w tym przypadku podstawić na metodę Cloppera-Pearsona, którą cechowało wysokie prawdopodobieństwo dla dużych n , a nawet skrajnych wartości p . Metoda Wilsona sprawdzała się nie najgorzej i być może – ze względu na fakt, że przedziały ufności są krótkie – to ona stanowi złoty środek w doborze najoptymalniejszej z metod.

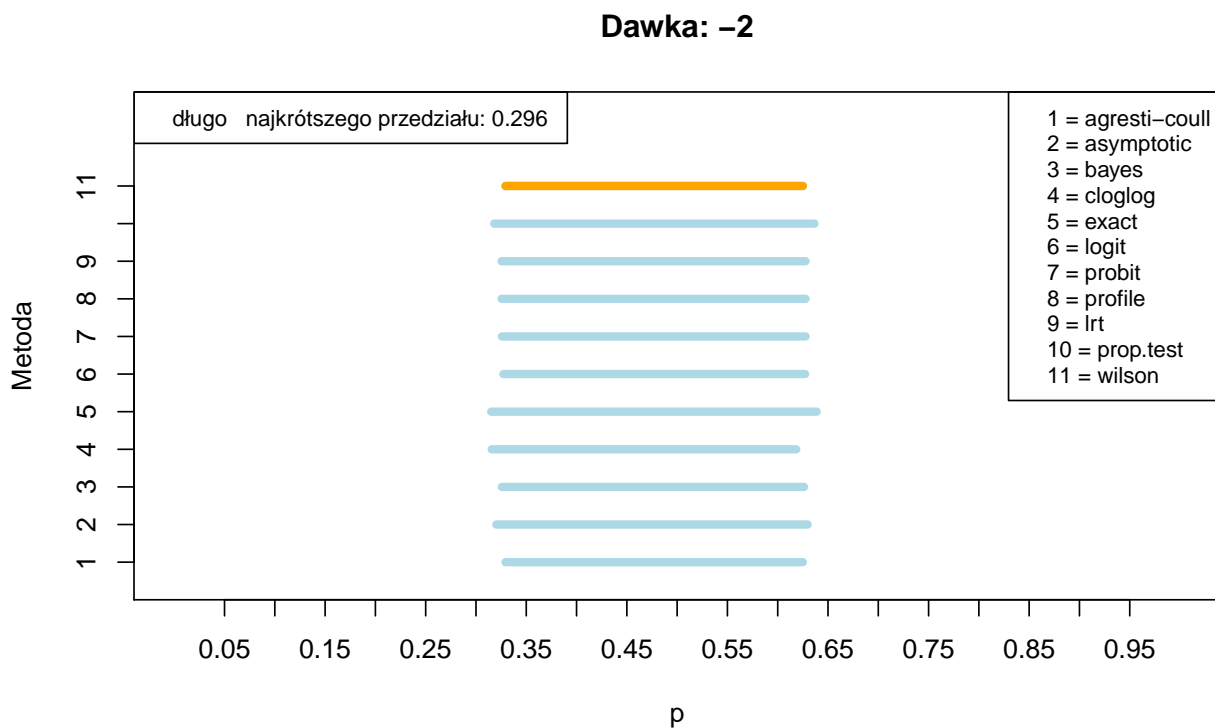
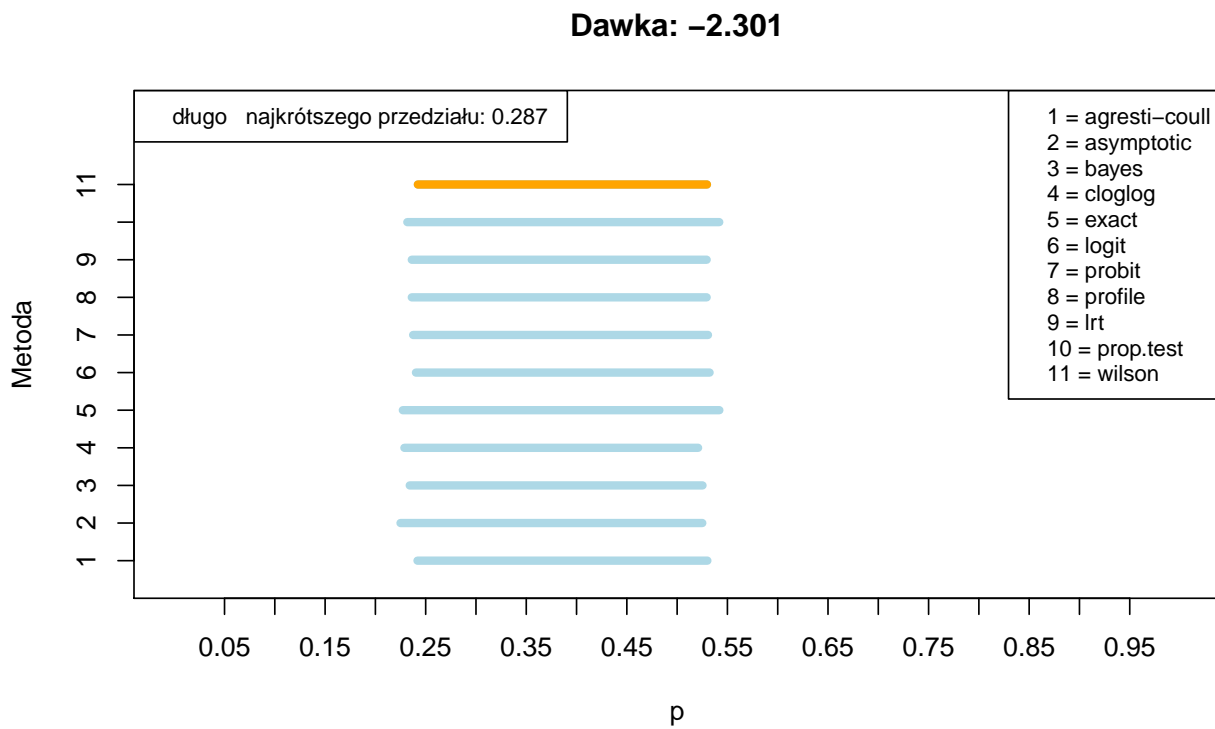
Dawka: -2.903



Dawka: -2.602



Rysunek 4: Porównanie wszystkich metod



Rysunek 5: Porównanie wszystkich metod

3 Lista 4 i 5

3.1 Zadanie 1

Przedstawimy teraz dwie funkcje z pakietu `stats`: `binom.test()` oraz `prop.test()`. Obie funkcje są odpowiedzialne za przeprowadzanie testu hipotezy prostej dotyczącej parametru p w rozkładzie dwumianowym $\mathcal{B}(1, p)$. Funkcja `binom.test()` przeprowadza test dokładny, z kolei funkcja `prop.test()` przeprowadza test asymptotyczny, co więcej, pozwala ona weryfikować hipotezy dotyczące wielu rozkładów dwumianowych. Obie te funkcje są jednak przydatne w testowaniu hipotez typu (dla ustalonej wartości p_0):

- $H_0 : p = p_0$ przeciwko $H_1 : p \neq p_0$,
- $H_0 : p \geq p_0$ przeciwko $H_1 : p < p_0$,
- $H_0 : p \leq p_0$ przeciwko $H_1 : p > p_0$.

3.2 Zadanie 2

Na podstawie danych zawartych w pliku `Reakcja.csv`, na poziomie istotności $\alpha = 0.05$ będziemy weryfikować różne hipotezy. W tym celu przeprowadzimy testy, o których mowa w Zadaniu 1.

- (a) *Prawdopodobieństwo poprawy stanu zdrowia pacjenta leczonego w domu najmniejszą dawką leku jest mniejsze bądź równe $1/2$.*

Naszą najmniejszą dawką leku jest wartość -2 . Za pomocą funkcji `filter()` wybierzemy z naszych danych tylko te dotyczące pacjentów leczonych w domu, po czym przeprowadzimy test.

```
zad2 %>% filter(Miejsce == "Dom", Reakcja == "Tak") %>% count(Dawka)
```

```
binom.test(x=6, n=14, p=0.5, alternative = "greater", conf.level = 0.95)

##
## Exact binomial test
##
## data: 6 and 14
## number of successes = 6, number of trials = 14, p-value = 0.788
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
## 0.2060733 1.0000000
## sample estimates:
## probability of success
## 0.4285714
```

Widzimy, że estymowane prawdopodobieństwo poprawy stanu zdrowia pacjenta leczonego w domu najmniejszą dawką leku jest równe około 0.429, czyli zachodzi nierówność $p = 0.428 \leq 0.5$. Co więcej, wartość p -value to $0.788 > \alpha$, zatem nie mamy podstaw do odrzucenia hipotezy.

- (b) *Prawdopodobieństwo poprawy stanu zdrowia pacjenta leczonego w domu najmniejszą dawką leku jest równe prawdopodobieństwu poprawy stanu zdrowia pacjenta leczonego najmniejszą dawką leku, ale w szpitalu.*

Skorzystamy tutaj z funkcji `prop.test()`, ponieważ mamy do porównania parametry dwóch rozkładów dwumianowych.

```
prop.test(c(6,13), c(14, 39), alternative = "two.sided",
          conf.level = 0.95)

##
## 2-sample test for equality of proportions with continuity correction
##
## data:  c(6, 13) out of c(14, 39)
## X-squared = 0.097708, df = 1, p-value = 0.7546
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2517698  0.4422460
## sample estimates:
##      prop 1      prop 2
## 0.4285714 0.3333333
```

Estymowalne prawdopodobieństwo sukcesu, że pacjent leczony w szpitalu wyzdrowieje wynosi około $p = 0.33$. Dodatkowo widzimy, że wartość p -value jest równa $0.7546 > \alpha$, zatem nie mamy podstaw do odrzucenia hipotezy.

Powtórzymy badania, tym razem dla największej dawki leku, wynoszącej -3.204 .

- (a) *Prawdopodobieństwo poprawy stanu zdrowia pacjenta leczonego największą dawką leku jest mniejsze bądź równe 0.5.*

```
##      Dawka n
## 1      -2 6
## 2 -2,301 5
## 3 -2,602 2
## 4 -2,903 1
```

```
binom.test(x=0, n=14, p=0.5, alternative = "greater", conf.level = 0.95)

##
## Exact binomial test
##
## data:  0 and 14
## number of successes = 0, number of trials = 14, p-value = 1
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0 1
## sample estimates:
## probability of success
##                                0
```


Widzimy, że u żadnego z pacjentów leczonych w domu największa dawka leku nie nastąpiła poprawa. Zatem estymowane prawdopodobieństwo poprawy wynosi 0 przy p -value równym 1.

- (b) *Prawdopodobieństwo poprawy stanu zdrowia pacjenta największą dawką leku jest równe prawdopodobieństwu poprawy stanu zdrowia pacjenta leczonego największą dawką leku, ale w szpitalu.*

Na wstępie możemy podejrzewać, że nie zajdzie tutaj równość, ponieważ u pacjentów leczonych największą dawką w szpitalu zaobserwowano poprawę, lecz u tych leczonych w domu – nie. Korzystając z funkcji `prop.test` porównamy parametry rozkładów dwumianowych.

```
prop.test(c(0,3), c(14, 39), alternative = "two.sided",
          conf.level = 0.95)

##
## 2-sample test for equality of proportions with continuity correction
##
## data:  c(0, 3) out of c(14, 39)
## X-squared = 0.15547, df = 1, p-value = 0.6934
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.20908806  0.05524191
## sample estimates:
##      prop 1      prop 2
## 0.00000000 0.07692308
```

Estymowane prawdopodobieństwo poprawy w tym wypadku wyniosło 0.0769. Ta liczba jest różna od zera, jednak wartość krytyczna p -value wynosi $0.6934 > \alpha$, zatem mimo wszystko nie mamy podstaw do odrzucenia hipotezy.

3.3 Zadanie 3

Teraz, w celu porównania testów dokładnego i asymptotycznego, przeprowadzimy odpowiednie symulacje, dzięki którym zbadamy, jakie czynniki wpływają na dokładność wynonywanego testu. Dla rozkładu dwumianowego $\mathcal{B}(1, p)$ na poziomie istotności 0.05 będziemy testować hipotezę dwustronną

$$H_0 : p = p_0 \quad \text{przeciwko} \quad H_1 : p \neq p_0.$$

Uwzględnimy różne rozmiary prób oraz różne wartości alternatyw; skoncentrujemy się na wartościach $n \in \{10, 50, 100, 1000, 10000\}$ oraz $p_0 \in \{0.05, 0.25, 0.5\}$, przy czym dobraliśmy $p_0 \leq 0.5$ ze względu na symetrię.

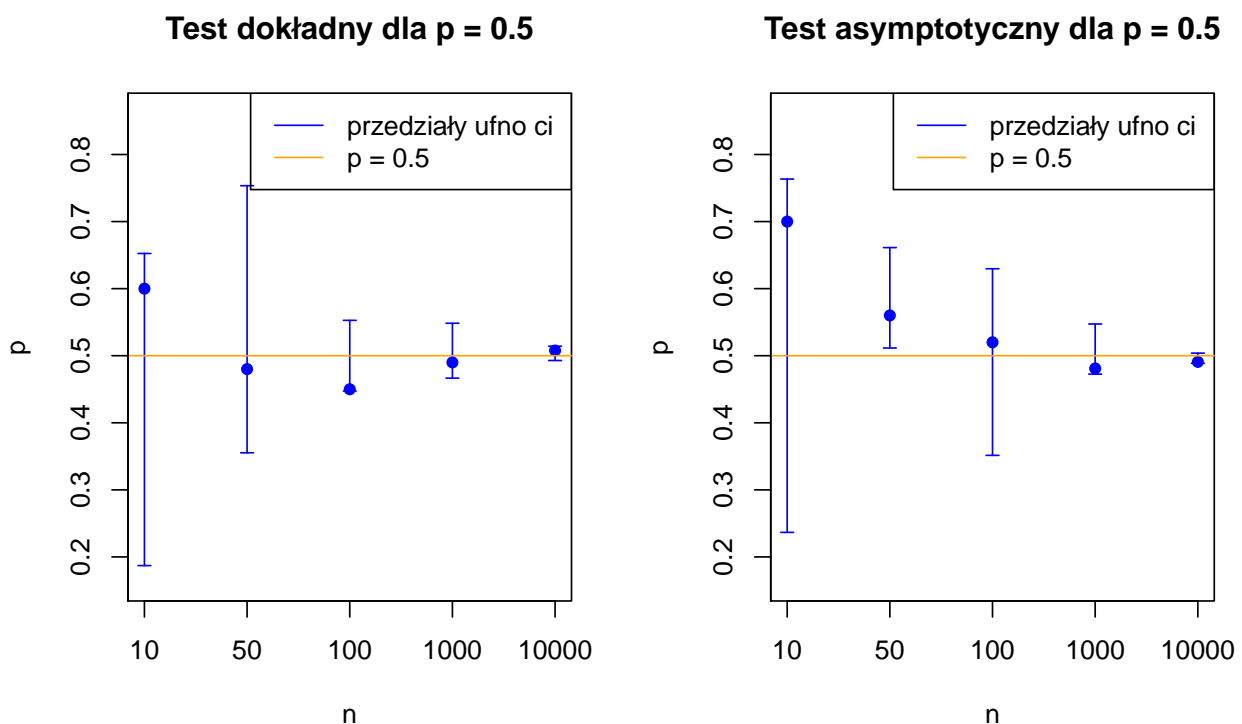
Zacniemy od przypadku dla $p_0 = 1/2$. Najpierw stworzymy funkcję, która będzie generowała dane do zestawiania, tzn. p -value, przedziały ufności oraz estymowane wartości parametru p . Poniżej prezentujemy fragment kodu.

```
simulation <- function(p0, alt = "two.sided", ex = TRUE){
  ns <- c(10, 50, 100, 1000, 10000)
```

```

p.vals <- c()
p.estims <- c()
l.confint <- c()
r.confint <- c()
for(i in 1:5){
  X <- rbinom(ns[i], 1, p0)
  suc <- length(which(X==1))
  if(ex == TRUE){
    p.test <- binom.test(c(suc, ns[i]-suc), ns[i], p = p0, alternative = alt)
  }
  else{
    p.test <- prop.test(suc, ns[i], p = p0, alternative = alt)
  }
  p.vals[i] <- p.test$p.value
  p.estims[i] <- p.test$estimate
  l.confint[i] <- p.test$conf.int[1]
  r.confint[i] <- p.test$conf.int[2]
}
return(data.frame(p.vals, p.estims, l.confint, r.confint))
}

```



Rysunek 6: Porównanie przedziałów ufności dla $p_0 = 0.5$

Obserwujemy, że długości przedziałów ufności maleją wraz ze wzrostem n . Dla $n = 10$ przedziały są nieefektywnie długie w obu przypadkach, a estymowane wartości p dalekie od

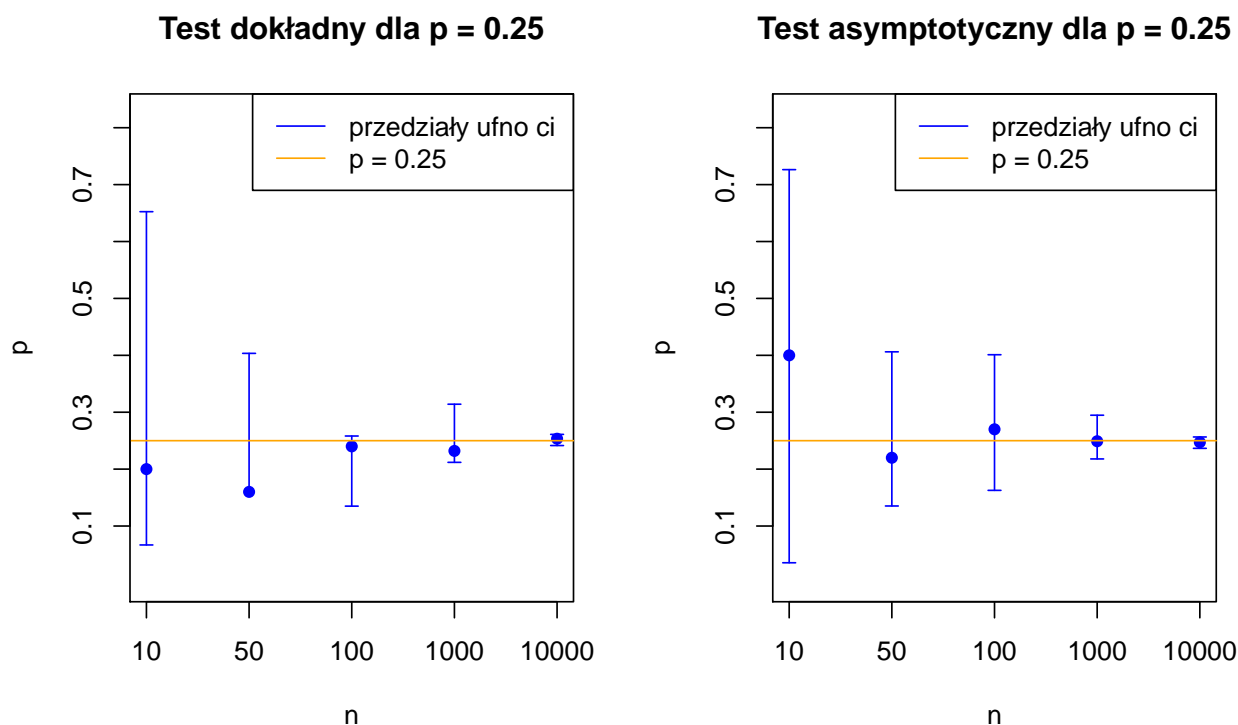
rzeczywistych, choć test dokładny sprawdza się w tym przypadku lepiej niż asymptotyczny. Ta tendencja jest na ogół (oprócz $n = 100$) powtarzalna. Dla $n \geq 1000$ mamy podobne wyniki dla obu testów, z niewielką przewagą testu dokładnego. Zauważmy też, że dla $n = 50$ przedziały ufności generowane przez test asymptotyczny nie zawierają punktu 0.5. Poniżej podajemy również odpowiednie wartości p -value dla obu testów.

	test dokładny	test asymptotyczny
$n = 10$	0.754	0.752
$n = 50$	0.888	0.322
$n = 100$	0.764	1.000
$n = 1000$	0.155	0.393
$n = 10000$	0.787	0.976

Tabela 9: Porównanie p -value dla $p_0 = 0.5$

Widzimy, że dla małych wartości n pod względem p -value oba testy wypadają podobnie, z minimalną przewagą testu dokładnego, który przeważa znacznie dla $n = 50$. Jednak przy większych rozmiarach prób, tzn. dla $n \geq 100$, lepsze wyniki zdaje się osiągać test asymptotyczny, w szczególności dla $n = 100$.

Pozostałe przypadki zbadamy bardziej zwięźle. Weźmy teraz $p_0 = 0.25$. Badane testy generują nam następujące przedziały ufności.



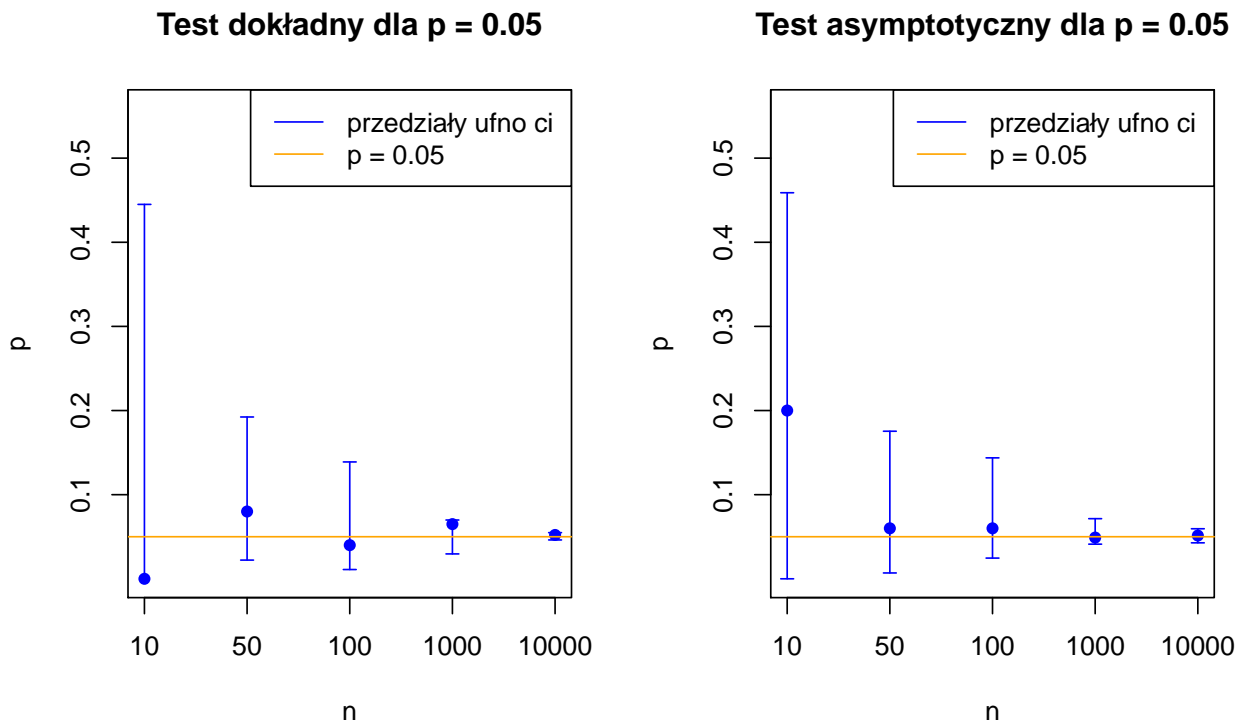
Rysunek 7: Porównanie przedziałów ufności dla $p_0 = 0.25$

	test dokładny	test asymptotyczny
$n = 10$	0.718	1.000
$n = 50$	0.745	0.191
$n = 100$	0.419	0.564
$n = 1000$	0.884	0.742
$n = 10000$	0.077	0.603

Tabela 10: Porównanie p -value dla $p_0 = 0.25$

W tym przypadku test dokładny zdaje się konstruować krótsze przedziały ufności oraz lepiej przybliżać rzeczywistą wartość p_0 (z wyjątkiem $n = 50$, gdzie lepszy jest test asymptotyczny). Jeśli spojrzymy zaś na wartości p -value, widzimy, że pod tym względem test asymptotyczny lepiej radzi sobie dla skrajnych wartości n .

Spróbujmy teraz wykonać symulacje dla skrajnie małych p . Dla $p = 0.05$ odpowiednie wykresy prezentują się następująco.

Rysunek 8: Porównanie przedziałów ufności dla $p_0 = 0.05$

	test dokładny	test asymptotyczny
$n = 10$	0.401	1.000
$n = 50$	0.181	0.516
$n = 100$	0.492	1.000
$n = 1000$	0.468	0.717
$n = 10000$	0.251	0.136

Tabela 11: Porównanie p -value dla $p_0 = 0.05$

Jak się okazuje, dla skrajnych wartości p_0 oba testy dają przedziały ufności zbliżonej długości, lecz to test asymptotyczny zdaje się osiągać lepsze wyniki – zarówno jeśli chodzi o estymację wartości p_0 , jak i p -values, które są większe dla prawie wszystkich wartości n . Jedyne w przypadku bardzo dużego rozmiaru próby to test dokładny daje większą (choć i tak stosunkowo małą) p -value.

Wniosek. Odpowiedź na pytanie, który z badanych testów jest lepszy, jak to zazwyczaj bywa, nie jest jednoznaczna. Jeśli estymowane prawdopodobieństwo sukcesu jest bliskie wartości 0.5 dla prób małego ($n < 100$) rozmiaru lepiej wybrać test dokładny, zaś dla dużego – test asymptotyczny. Jeśli estymujemy skrajną wartość prawdopodobieństwa sukcesu, rozsądnym wyborem wydaje się być test asymptotyczny, chyba że mamy do czynienia z bardzo małą próbą. Wówczas to test dokładny sprawdzi się lepiej.