Analiza danych ankietowych

Sprawozdanie 1

Weronika Jaszkiewicz Weronika Pyrtak

Spis treści

1 Część I

1.1 Zadanie 1

W pewnej dużej firmie technologicznej przeprowadzono ankietę mającą na celu ocenę skuteczności programów szkoleniowych dla pracowników. Wzięło w niej udział 200 losowo wybranych osób (losowanie proste ze zwracaniem).

1.1.1 Zadanie 1.1

Wczytywanie danych z pliku ankieta.csv.

Powyższe dane zawierają 200 wierszy oraz 8 kolumn.

Następnie sprawdzono typy przyjmowanych zmiennych.

Tabela 1: Typy zmiennych

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

Zamieniono zmienne o typie *character* na typ *factor*.

Przeszukano zbiór pod względem braków danych.

Liczba wartości brakujących wynosi: 0

Sprawdzono, czy typy zmiennych zostały prawidłowo rozpoznane.

1. zmienne ilościowe (typ numeric)

Tabela 2: Zmienne ilościowe

STAŻ	2
PYT_1	4
PYT_2	5
PYT_3	6
WIEK	8

Liczba zmiennych ilościowych: 5

2. zmienne jakościowe (typ factor)

Tabela 3: Zmienne jakościowe (factor)

DZIAŁ	1
CZY_KIER	3
PŁEĆ	7

Liczba zmiennych jakościowych (typ factor): 3

1.1.2 zadanie 1.2

Utworzono zmienna “WIEK_KAT” przeprowadzająx kategoryzację zmiennej “WIEK” korzystając z następujących przedziałów do 35 lat, mi¸edzy 36 a 45 lat, mi¸edzy 46 a 55 lat, powyżej 55 lat.

1.1.3 zadanie 1.3

Sporządzono tablice liczności dla zmiennych: DZIAŁ, STAŻ, CZY_KIER, PŁEĆ, WIEK_KAT.

Tabela 4: Tablica ilości dla zmiennej DZIAŁ

	HR	IT	MK	PD
Ilość	31	26	45	98

Tabela 5: Tablica ilości dla zmiennej STAŻ

	<1 rok	1-2 lata	3+ lat
Ilość	41	140	19

Tabela 6: Tablica ilości dla zmiennej CZY KIER

	TAK	NIE
Ilość	173	27

Tabela 7: Tablica ilości dla zmiennej PŁEĆ

	Kobieta	Mężczyzna
Ilość	71	129

Tabela 8: Tablica ilości dla zmiennej WIEK KAT

	0-35	36-45	46-55	55+
Ilość	26	104	45	25

WNIOSKI:

Największą grupę pracowników stanowi dział PD (98 osób), a najmniejszą IT (26 osób). Działy HR i MK mają pośrednie wartości (odpowiednio 31 i 45 osób). Może to wskazywać na różne zapotrzebowanie na pracowników w poszczególnych działach.

Większość pracowników (140 osób) ma od 1 do 2 lat stażu. 41 osób pracuje krócej niż rok, a tylko 19 osób ma staż powyżej 3 lat. Wskazuje to na dużą rotację pracowników lub na to, że firma stosunkowo niedawno zatrudniła większość obecnej kadry.

Większość pracowników (173 osoby) pełni funkcje kierownicze, a jedynie 27 osób nie. Może to oznaczać, że w firmie jest dużo stanowisk kierowniczych lub że definicja „kierownika" obejmuje szerokie spektrum stanowisk.

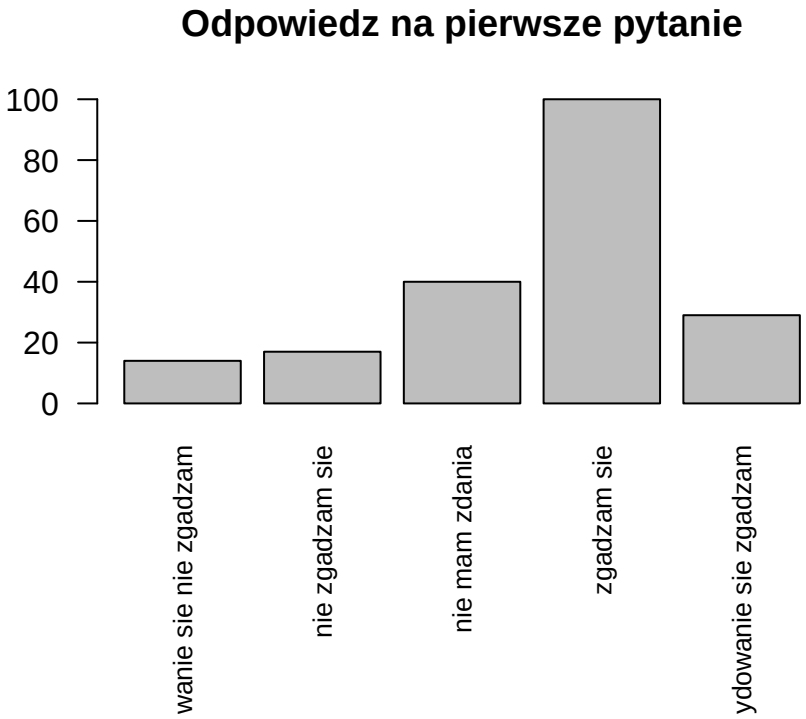
W firmie przeważają mężczyźni (129 osób) nad kobietami (71 osób). Może to wynikać z charakteru działalności firmy lub preferencji rekrutacyjnych.

Najwięcej pracowników jest w grupie wiekowej 36-45 lat (104 osoby). Pozostałe grupy wiekowe mają znacznie mniejszą reprezentację.

5 zmiennych ilościowych (np. staż, wiek, odpowiedzi na pytania PYT_1–PYT_3). 4 zmienne jakościowe, np. dział, płeć, kategoria wieku. Brak wartości brakujących wskazuje na kompletność i dobrą jakość danych.

1.1.4 Zadanie 1.4

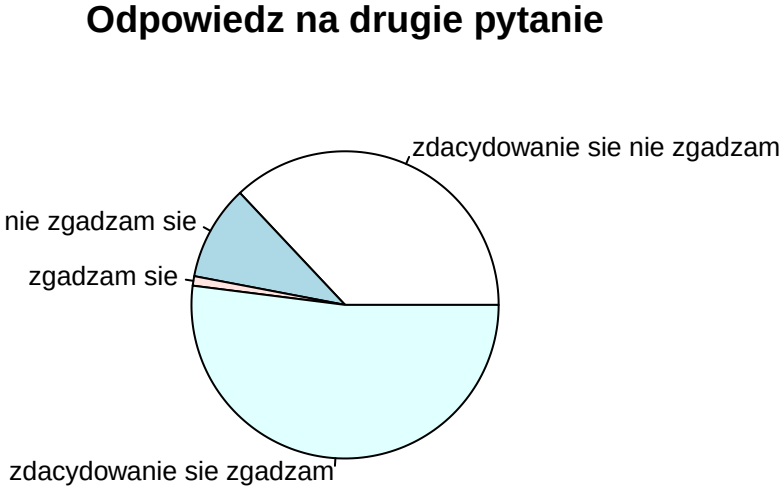
Sporządzono wykresy kołowe oraz wykresy słupkowe dla zmiennych: PYT_1 oraz PYT_2.

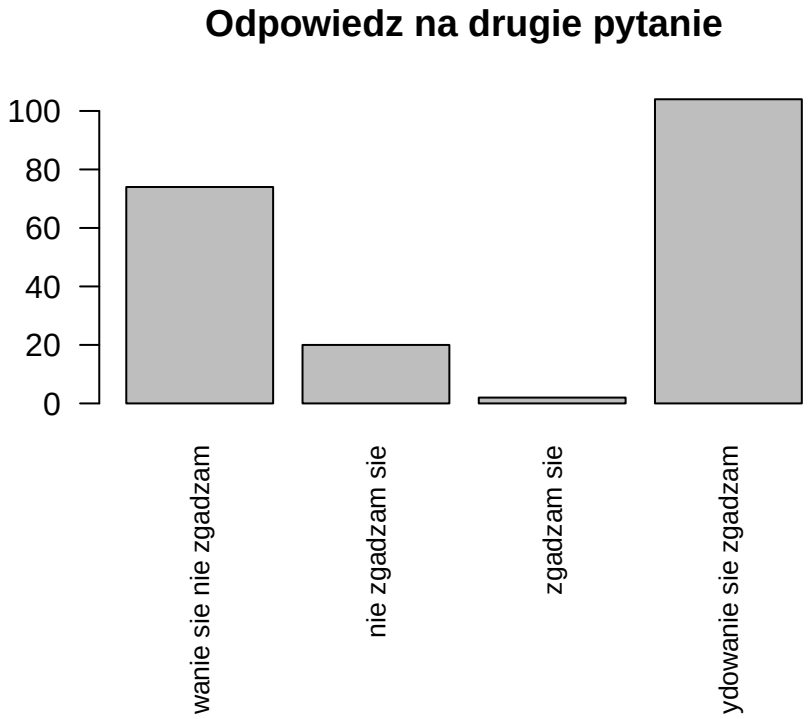


Największa liczba respondentów zaznaczyła odpowiedź "zgadzam się", co wskazuje, że ogólna ocena wsparcia i materiałów dostarczanych przez firmę jest pozytywna. Istnieje także pewna grupa osób, które "zdecydowanie się zgadzają", co podkreśla, że część pracowników uważa wsparcie za bardzo dobre.

Znaczna część pracowników wybrała odpowiedź "nie mam zdania", co może sugerować, że nie mieli oni okazji skorzystać ze wsparcia lub materiały nie są dla nich wystarczająco widoczne.

Pewna część respondentów zaznaczyła opcje "nie zgadzam się" i "zdecydowanie się nie zgadzam", ale ich liczba jest stosunkowo mała w porównaniu do grupy zadowolonych pracowników. Może to wskazywać na pewne niedociągnięcia w dostępie do materiałów lub ich jakości, jednak nie jest to powszechny problem.





Największa część respondentów zaznaczyła odpowiedź "zdecydowanie się zgadzam", co wskazuje, że większość pracowników uważa szkolenia oferowane przez firmę za dobrze dopasowane do ich potrzeb. Istnieje także spora grupa osób, które wybrały opcję "zgadzam się", co dodatkowo potwierdza ogólnie pozytywne nastawienie wobec polityki szkoleniowej firmy.

Stosunkowo niewielka część respondentów wybrała odpowiedź "nie zgadzam się" oraz "zdecydowanie się nie zgadzam", co sugeruje, że niektóre osoby mogą mieć trudności z dostępem do odpowiednich szkoleń lub nie uważają ich za skuteczne w kontekście swojego rozwoju zawodowego.

Pomimo że liczba negatywnych odpowiedzi jest niewielka, warto przeanalizować, czy istnieją konkretne obszary, w których szkolenia mogą być bardziej dostosowane do indywidualnych potrzeb

1.1.5 zadanie 1.5

Sporządzono 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.

Tabela 9: Tabela wielodzielcza dla zmiennych PYT 1 i DZIAL

	HR	IT	MK	PD
zdacydowanie się nie zgadzam	2	0	3	9
nie zgadzam się	2	2	3	10
nie mam zdania	5	4	14	17
zgadzam się	19	15	15	51
zdacydowanie się zgadzam	3	5	10	11

Najbardziej pozytywne opinie pochodzą z działów HR i PD – większość respondentów z tych działów wybrała opcje "zgadzam się" i "zdecydowanie się zgadzam". Dział MK (Marketing) jest bardziej podzielony – występuje większy odsetek osób, które nie mają zdania, co może sugerować, że dla tej grupy pytanie nie było jednoznaczne lub temat ich mniej dotyczył. Dział IT wykazuje najmniejszy poziom negatywnych odpowiedzi ("zdecydowanie się nie zgadzam" = 0), ale też stosunkowo niewiele osób udzieliło odpowiedzi skrajnie pozytywnej.

Tabela 10: Tabela wielodzielcza dla zmiennych PYT 1 i STAŻ

	<1 rok	1-2 lata	3+ lat
zdacydowanie się nie zgadzam	5	5	4
nie zgadzam się	6	10	1
nie mam zdania	8	26	6
zgadzam się	19	75	6
zdacydowanie się zgadzam	3	24	2

Najbardziej pozytywne odpowiedzi (zgadzam się i zdecydowanie się zgadzam) pochodzą od osób z doświadczeniem 1-2 lata, co sugeruje, że osoby na tym etapie kariery widzą największą wartość badanego zagadnienia. Osoby z najmniejszym stażem (<1 rok) są bardziej podzielone, częściej nie mają zdania lub udzielają odpowiedzi negatywnych. Osoby z największym stażem (3+ lata) rzadko wybierają odpowiedzi skrajne, co może oznaczać, że ich ocena sytuacji jest bardziej neutralna.

Pracownicy z 1-2 latami stażu są najbardziej pozytywnie nastawieni do badanego zagadnienia. Osoby z krótszym stażem mogą wymagać większej ilości informacji lub wsparcia, aby mogły bardziej świadomie ocenić sytuację.

Tabela 11: Tabela wielodzielcza dla zmiennych PYT 1 i CZY_KIER

	TAK	NIE
zdacydowanie się nie zgadzam	10	4
nie zgadzam się	14	3
nie mam zdania	34	6
zgadzam się	88	12
zdacydowanie się zgadzam	27	2

Osoby na stanowiskach kierowniczych częściej wybierają opcję "zgadzam się" i "zdecydowanie się zgadzam", co może świadczyć o ich większym zadowoleniu. Osoby niepełniące funkcji kierowniczych są bardziej podzielone – widoczny jest większy odsetek neutralnych i negatywnych odpowiedzi.

Pracownicy na stanowiskach kierowniczych mają bardziej pozytywne podejście do badanego aspektu, natomiast pracownicy bez funkcji kierowniczych mogą czuć się mniej związani z tematem lub mieć inne doświadczenia.

Mężczyźni są bardziej podzieleni, częściej wybierają opcje neutralne i negatywne. Kobiety częściej wybierają pozytywne odpowiedzi, co może świadczyć o lepszym dopasowaniu badanego zagadnienia do ich oczekiwań.

Tabela 12: Tabela wielodzielcza dla zmiennych PYT 1 i PŁEĆ

	Kobieta	Mężczyzna
zdacydowanie się nie zgadzam	3	11
nie zgadzam się	7	10
nie mam zdania	14	26
zgadzam się	36	64
zdacydowanie się zgadzam	11	18

Istnieją różnice w postrzeganiu badanego tematu w zależności od płci – warto byłoby zbadać, co wpływa na większą satysfakcję kobiet.

Tabela 13: Tabela wielodzielcza dla zmiennych PYT 1 i WIEK KAT

	0-35	36-45	46-55	55+
zdacydowanie się nie zgadzam	1	11	2	0
nie zgadzam się	6	7	1	3
nie mam zdania	3	24	5	8
zgadzam się	13	50	25	12
zdacydowanie się zgadzam	3	12	12	2

Najbardziej pozytywne odpowiedzi pochodzą od osób w wieku 36-45 lat oraz 46-55 lat – te grupy częściej wybierają opcje "zgadzam się" i "zdecydowanie się zgadzam". Najmłodsza grupa (0-35 lat) oraz najstarsza grupa (55+) mają wyższy odsetek odpowiedzi neutralnych lub negatywnych.

Pracownicy w średnim wieku są najbardziej pozytywnie nastawieni do badanego zagadnienia. Osoby młodsze i starsze mogą mieć inne oczekiwania lub mniej doświadczenia w tym obszarze.

1.1.6 zadanie 1.6

Sporządzono tablicę wielodzielczą dla pary zmiennych: PYT_2 i PYT_3.

Tabela 14: Tabela wielodzielcza dla zmiennych PYT 1 i PYT 2

	zdacydowanie się nie zgadzam	nie zgadzam się	zgadzam się	zdacydo
zdacydowanie się nie zgadzam	13	0	1	
nie zgadzam się	16	0	0	
nie mam zdania	39	0	1	
zgadzam się	3	17	0	
zdacydowanie się zgadzam	3	3	0	

Dane obejmują zarówno zmienne ilościowe (np. STAŻ, WIEK, PYT_1), jak i jakościowe (np. DZIAŁ, PŁEĆ, CZY_KIER). Brak wartości brakujących sugeruje, że zestaw danych jest kompletny i dobrze przygotowany do analizy.

W zbiorze znajduje się 5 zmiennych ilościowych i 4 zmienne jakościowe (faktory).

Najwięcej osób w próbie należy do działu PD (98 osób), najmniej do IT (26 osób).

Większość badanych ma staż pracy 1-2 lata (140 osób), co może wskazywać na młodą kadrę pracowników.

Większość respondentów pełni funkcję kierowniczą (173 TAK vs. 27 NIE).

Przewaga mężczyzn w próbie (129 mężczyzn vs. 71 kobiet).

Największa grupa wiekowa to 36-45 lat (104 osoby).

PYT_1 a DZIAŁ: Najwięcej pozytywnych odpowiedzi ("zgadzam się" lub "zdecydowanie się zgadzam") udzieliły osoby z działu PD.

PYT_1 a STAŻ: Wśród osób z najkrótszym stażem (<1 rok) więcej było odpowiedzi neutralnych lub negatywnych. Może to wynikać z braku doświadczenia lub innego spojrzenia na temat badania.

PYT_1 a CZY_KIER: Kierownicy częściej zgadzali się z twierdzeniem zawartym w PYT_1 niż osoby niepełniące funkcji kierowniczych.

PYT_1 a PŁEĆ: Mężczyźni częściej wyrażali zdecydowaną opinię niż kobiety, które częściej wybierały opcję "nie mam zdania".

PYT_1 a WIEK_KAT: Najwięcej osób w wieku 36-45 lat zgadzało się z PYT_1, natomiast najmniej zdecydowanych odpowiedzi było w grupie 0-35 lat.

PYT_1 a PYT_2: Istnieje silna zależność – osoby, które nie zgadzały się z PYT_1, często miały podobne odpowiedzi na PYT_2, a ci, którzy zgadzali się z PYT_1, także zgadzali się z PYT_2.

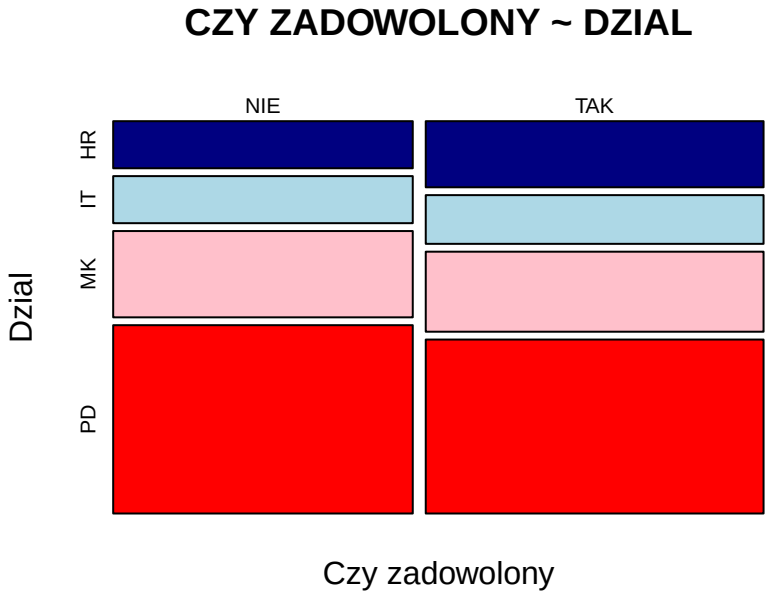
Występują różnice w opiniach w zależności od działu, stażu pracy, płci, funkcji kierowniczej i wieku. Kierownicy oraz osoby z dłuższym stażem mają bardziej zdecydowane opinie. Istnieje silna korelacja między odpowiedziami na PYT_1 i PYT_2. Możliwe różnice w podejściu do badanej kwestii między kobietami a mężczyznami oraz młodszymi a starszymi pracownikami.

1.1.7 zadanie 1.7

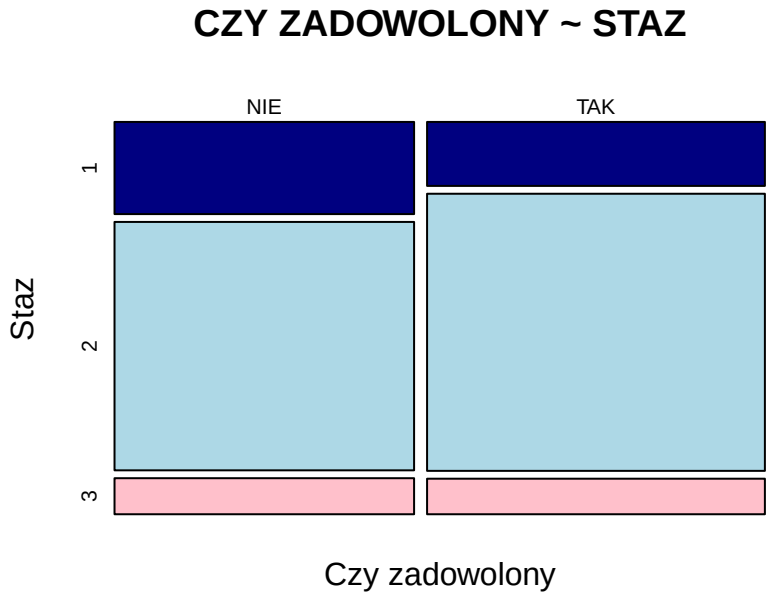
Utworzono zmienną CZY_ZADOW na podstawie zmiennej PYT_2 ł ¸acz ¸ac kategorie "nie zgadzam się" i "zdecydowanie się nie zgadzam" oraz "zgadzam się" i "zdecydowanie się zgadzam"

1.1.8 zadanie 1.8

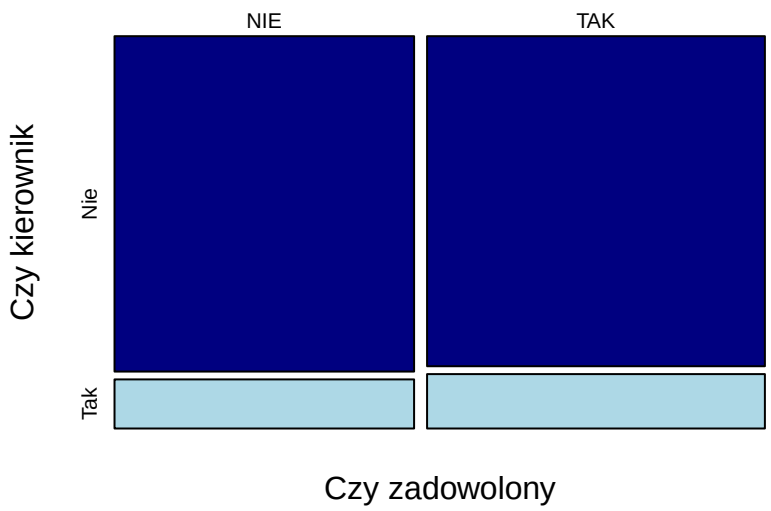
Sporządzono wykresy mozaikowe odpowiadające parom zmiennych CZY_ZADOW i DZIAŁ, CZY_ZADOW i STAŻ, CZY_ZADOW i CZY_KIER, CZY_ZADOW i PŁEĆ oraz CZY_ZADOW i WIEK_KAT.


CZY ZADOWOLONY ~ DZIAL
NIE
TAK
HR
IT
MK
PD
Dzial
Czy zadowolony

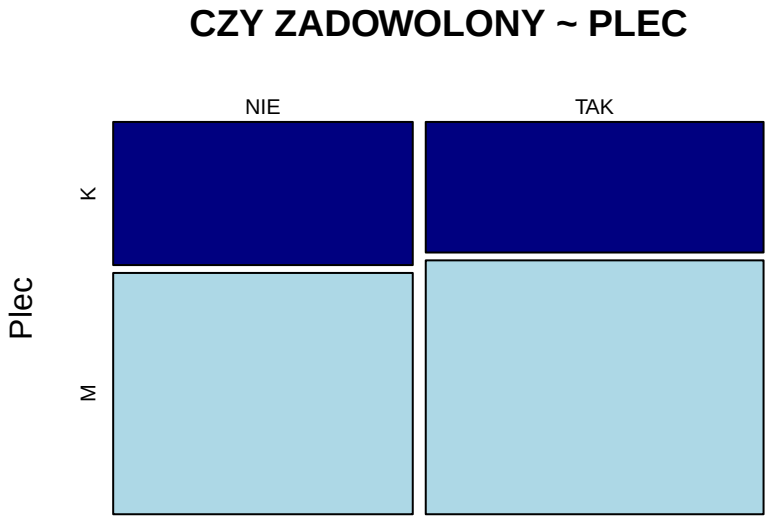
```
## NULL
```


CZY ZADOWOLONY ~ STAZ
NIE
TAK
1
2
3
Staz
Czy zadowolony

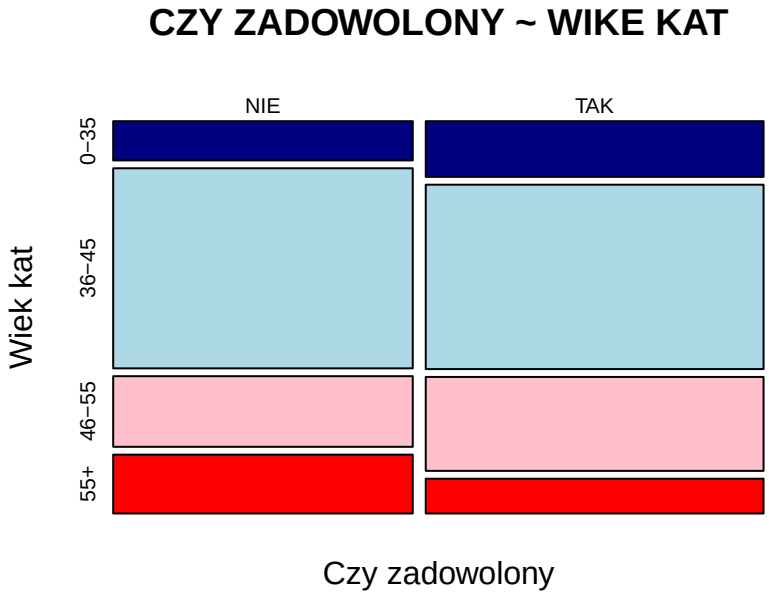
```
## NULL
```


CZY ZADOWOLONY ~ CZY KIER
NIE
TAK
Czy kierownik
Nie
Tak
Czy zadowolony

```
## NULL
```


CZY ZADOWOLONY ~ PLEC
NIE
TAK
Plec
K
M
Czy zadowolony

```
## NULL
```

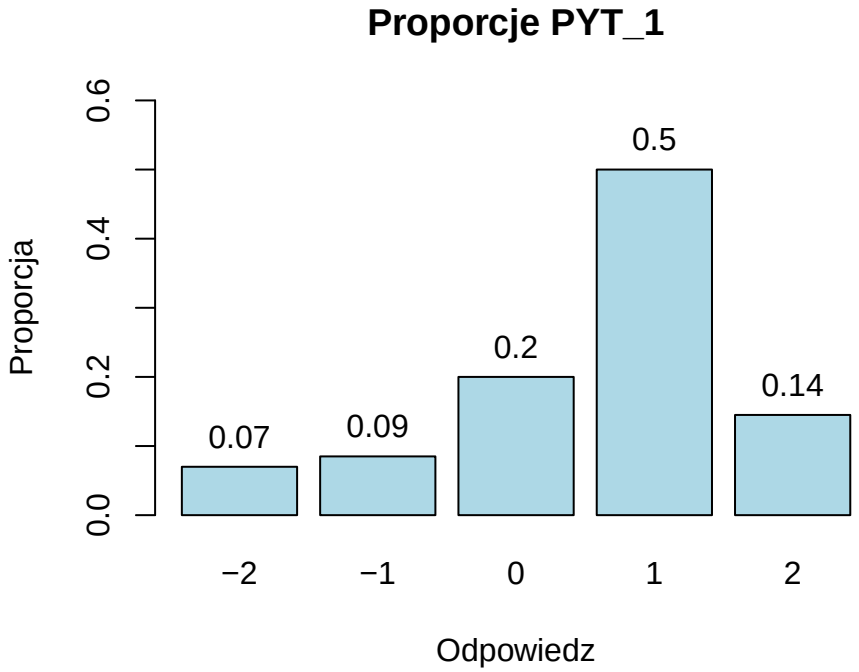



```
## NULL
```

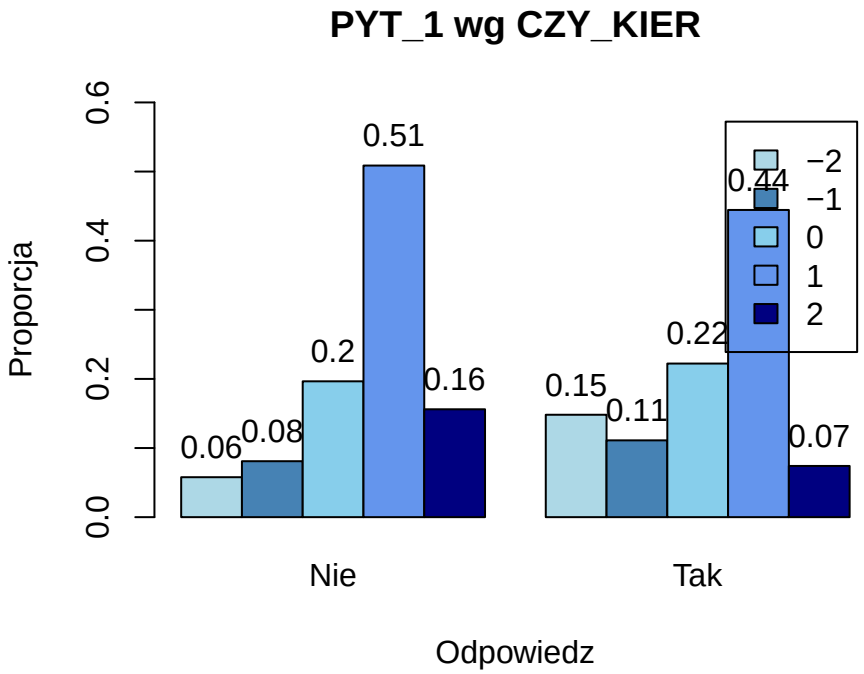
2 Część II

2.1 Zadanie 2

Na wykresie słupkowym została przedstawiona proporcja odpiedzi pracowników firmy na pytanie *PYT_1: "Jak bardzo zgadzasz się ze stwierdzeniem, że firma zapewnia odpowiednie wsparcie i materiały umożliwiaja¸ce skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń?"*. Z wykresu wynika, że połowa badanych zgadza się ze stwierdzeniem, że firma zapewnia odpowiednie wsparcie i materialy umożliwiające skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń, ponadto 14% badanych zdecydowanie popiera tę tezę, a 20% nie ma zdania na ten temat. Natomiast niecałe 10% nie zgadza się z powyższym stwierdzeniem, a 7% uważa, że jest ono zdecydowanie sprzeczne.



Ponadto sprawdzono jak rozkłada się poziom zgodności z powyższym stwierdzeniem względem pełnionego stanowiska (kierownicze lub niekierownicze) dzięki zmiennej CZY_KIER. Z wykresu można wywyniksować, że w obu przypadkach około połowy badanych zgadza się ze stwierdzeniem z PYT_1, jednak ponad dwókrotnie większa część osób (procentowo) bez stanowiska kierowniczego niż na stanowisku kierowniczym jest zdecydowanie zadowolona z udostępnianych materiałów ze szkoleń. Również można zauważyć, że odpowiedzi *nie mam zdania/ nie zgadzam się/ zdecydowanie się nie zgadzam* zanaczyło większy odetek osób na stanowistach kierowniczych niż nie. Zatem z analizy wykresu pudełkowego wynika, że pracownicy na stanowiskach kierowniczych są mniej zdowoleni ze wsparcia i materiałów zapewnianych przez firmę umożliwiaja¸ce skuteczne wykorzystanie w praktyce wiedzy zdobytej w trakcie szkoleń.



2.2 Zadanie 3

Funkcja *sample()* z biblioteki *stats* losuje próbkę z podanego zbioru danych. **Składnia:** sample(x, size, replace, prob) Gdzie: - x - wektor do losowania - size - liczba elementów do wylosowania - replace - określa czy losowanie jest ze zwracaniem (TRUE/FALSE) - prob - prawdopodobieństwa dla poszczególnych elementów(parametr opcjonalny) **Przykłady użycia**

```
## [1] 6 1 3 2 9
```

```
## [1] 1 3 9 3 6
```

```
## [1] 5 4 8 1 9
```

Następnie z rekordów zawartych w pliku ankieta.csv zostało wylosowane 10% losowych ze wszystkich rekordów za pomocą losowania ze zwracaniem oraz bez zwracania. **Losowanie wierszy ze zwracaniem**

```
##      DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK WIEK_KAT CZY_ZADOW
## 127     MK    2      Nie     1     2     2    K   36    36-45       TAK
## 45      PD    1      Nie     1     2     2    M   36    36-45       TAK
## 146     MK    3      Nie     1     2     2    K   52    46-55       TAK
## 77      PD    2      Nie     1     2     1    K   41    36-45       TAK
## 176     HR    2      Nie     1    -1    -1    M   40    36-45       NIE
## 17      IT    2      Nie     0    -2    -2    K   45    36-45       NIE
## 6       IT    3      Tak     0     1     1    K   57      55+       TAK
## 138     MK    2      Nie     2     2     2    K   39    36-45       TAK
## 87      PD    1      Nie    -2    -2    -2    M   49    46-55       NIE
## 86      PD    2      Tak     1     2     2    M   52    46-55       TAK
```

```
## 64      PD    2      Nie     2    -1    -1    M   53    46-55       NIE
## 33      PD    1      Nie     1     2     2    M   30     0-35       TAK
## 32      PD    1      Nie     0    -2    -2    M   33     0-35       NIE
## 60      PD    2      Nie     1    -1     1    M   55    46-55       NIE
## 39      PD    1      Nie     1     2     2    M   40    36-45       TAK
## 5       IT    3      Tak     1     2    -1    K   65      55+       TAK
## 196     HR    2      Nie     1     2     2    M   42    36-45       TAK
## 74      PD    2      Nie     2     2     2    K   44    36-45       TAK
## 198     HR    2      Nie    -1    -2    -2    K   39    36-45       NIE
## 77.1    PD    2      Nie     1     2     1    K   41    36-45       TAK
```

Losowanie wierszy bez zwracania

```
##     DZIAŁ STAŻ CZY_KIER PYT_1 PYT_2 PYT_3 PŁEĆ WIEK WIEK_KAT CZY_ZADOW
## 90     PD    1      Nie     1     2     2    M   53    46-55       TAK
## 182    HR    2      Nie     1     2     2    M   40    36-45       TAK
## 83     PD    1      Nie     1    -2    -1    M   52    46-55       NIE
## 156    MK    2      Nie     0    -2    -2    M   37    36-45       NIE
## 98     PD    2      Nie     0    -2    -2    M   40    36-45       NIE
## 128    MK    2      Nie     0    -2    -2    K   45    36-45       NIE
## 138    MK    2      Nie     2     2     2    K   39    36-45       TAK
## 35     PD    1      Nie    -1    -2    -2    M   28     0-35       NIE
## 13     IT    2      Tak     1     2     2    K   48    46-55       TAK
## 18     IT    2      Nie     1     2     2    K   43    36-45       TAK
## 173    HR    2      Tak     1     2     1    M   39    36-45       TAK
## 136    MK    2      Nie    -2    -2     2    K   42    36-45       NIE
## 37     PD    1      Nie     2     2     2    M   33     0-35       TAK
## 108    PD    2      Tak    -2    -2    -2    K   40    36-45       NIE
## 145    MK    3      Nie     0    -2    -2    K   46    46-55       NIE
## 38     PD    1      Nie     0    -2    -1    M   35     0-35       NIE
## 140    MK    2      Nie     2     2    -1    K   38    36-45       TAK
## 132    MK    3      Tak     0    -2    -2    K   42    36-45       NIE
## 174    HR    2      Nie     1     2     1    M   36    36-45       TAK
## 93     PD    2      Nie     2     2     2    M   37    36-45       TAK
```

2.3 Zadanie 4

Funkcja *binomial_sim* dla każdej próbki generuje 1 lub 0 z prawdopodobieństwem p i zwraca realizację próby. Funckja przyjmuje dwa parametry: - n - długość próby - p - teoretyczne prawdopodobieństwo sukcesu.

Funkcja *binomial___N_sim* wykonuje N prób Monte Carlo z rozkładu dwumianowego korzystając z funkcji *binomial_sim*, zwraca wketor realizacji prób z rozkładu dwumianowego. Funkcja przyjmuje trzy paramatry: - n - długość próby - p - teoretyczne prawdopodobieństwo sukcesu - N - liczbę prób Monte Carlo

W celu przetestowania poprawności zaproponowanych funkcji, przeprowadzono symulację,

której celem było wygenerowanie wektora zmiennych losowych i obliczenie ich charakterystyk, a następnie porównanie wyników empirycznych z teoretycznymi. **Parametry symulacji:** - n = 100 - p = 0.5 - N = 10000 **Charakterystyki rozkładu, które zostały uwzględnione:** 1. **Średnia rozkładu:** - **Empiryczna:** obliczona za pomocą funkcji *mean()* na wygenerowanych danych, - **Teoretyczna:** obliczona na podstawie wzoru np, gdzie n to liczba prób, p to prawdopodobieństwo sukcesu. 2.**Odchylenie standardowe rozkładu:** - **Empiryczne:** obliczone za pomocą funkcji *sd()* na wygenerowanych danych, - **Teroretyczne:** obliczone na podstawie wzoru

$$\sigma = \sqrt{np(1-p)}$$

, gdzie n to liczba prób, p to prawdopodobieństwo sukcesu. 3. **Hstogram częstości:** - **Empiryczne:** przedstawione za pomocą funkcji *hist()* na podstawie wygenerowanych danych - **Teoretyczne:** porównane z teoretycznymi wartościami prawdopodobieństwa sukcesu, obliczonymi za pomocą funkcji *dbinom()*

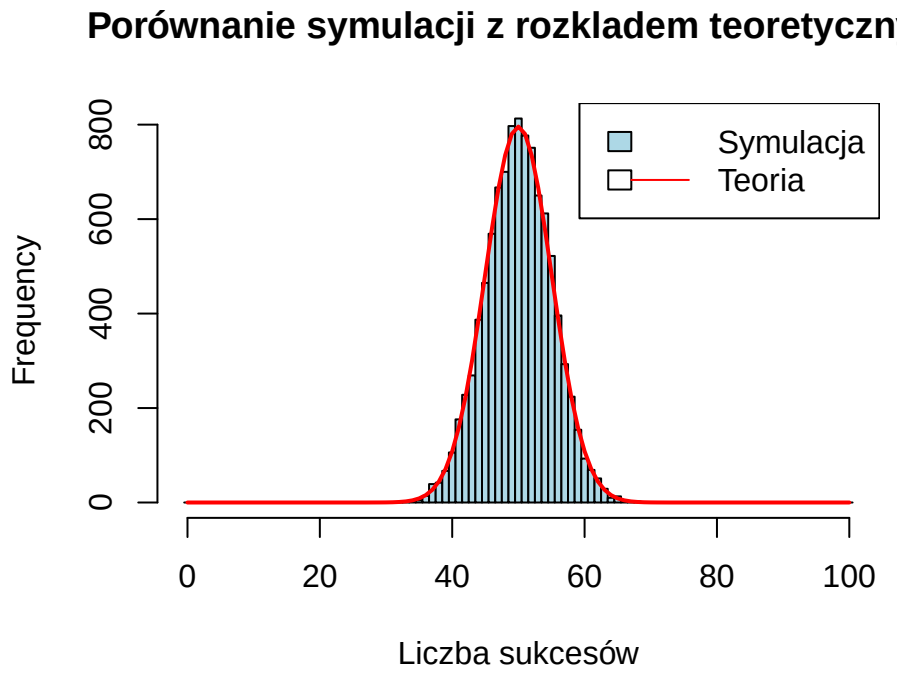
Wyniki symulacji sugerują, że funkcja poprawnie generuje zmienną losową z rozkładu dwumianowego.

```
## Teoretyczna średnia:  50
```

```
## Teoretyczne odchylenie standardowe:  5
```

```
## Empiryczna średnia:  50.0524
```

```
## Empiryczne odchylenie standardowe:  4.977362
```



2.4 Zadnia 5

Funkcja *wielomianowy_sim* generuje pojedynczą realizację rozkładu wielomianowego z prawdopodbieństwami sukcesu danymi jako wektor p. Przyjmuje dwa parametry: - n - liczba prób

(wielkość próby) - p - wektor prawdopodobieństw dla poszczególnych kategorii. Zwraca wektor liczności wystąpień każdej kategorii.

Funkcja *wielomianowy_N_sim* wykonuje N prób Monte Carlo dla rozkładu wielomianowego. Korzysta z funkcji *wielomianowy_sim*, aby wygenerować realizacje prób i zwraca macierz wyników. Przyjmuje trzy parametry: - n - liczba prób (wielkość próby) - p - wektor prawdopodobieństw dla poszczególnych kategorii - N - liczba prób Monte Carlo

Analogicznie do poprzedniego zadania, przeprowadzono symulację, której celem było sprawdzenie poprawności zaproponowanych funckji. Została wygenerowana macierz zawierająca realizaje zmiennych rozkładu wielomianowego, a następnie obliczono ich charakterystyki i porównano z teoretycznymi wartościami. **Parametry symulacji:** - n = 100 - p = [0.5, 0.1, 0.2, 0.2] - N = 10000 **Charakterystyki rozkładu, które zostały uwzględnione:** 1. **Średnia rozkładu:** - **Empiryczna:** obliczona za pomocą funkcji *rowMeans()* na wygenerowanych danych, - **Teoretyczna:** obliczona na podstawie wzoru np, gdzie n to liczba prób, p to prawdopodobieństwo sukcesu. 2.**Odchylenie standardowe rozkładu:** - **Empiryczne:** obliczone za pomocą funkcji *apply(simulated_data, 1, sd)* na wygenerowanych danych, - **Teroretyczne:** obliczone na podstawie wzoru

$$\sigma = \sqrt{np(1-p)}$$

, gdzie n to liczba prób, p to wektor prawdopodobińśtw.

Wyniki symulacji sugerują, że funkcja poprawnie generuje zmienną losową z rozkładu wielomianowego.

```
## Teoretyczna średnia:  50 10 20 20
```

```
## Teoretyczne odchylenie standardowe:  5 3 4 4
```

```
## Empiryczna średnia:  49.9756 10.05 20.0428 19.9316
```

```
## Empiryczne odchylenie standardowe:  4.996169 2.988811 3.989431 4.00149
```

3 Część III i IV

3.1 Zadanie 6

```
## Przedział ufności (Clopper-Pearson) dla 30 / 100 : 0.2124064 0.3998147
```

```
## Przedział ufności (Clopper-Pearson) dla danych binarnych: 0.1947936 0.378667
```

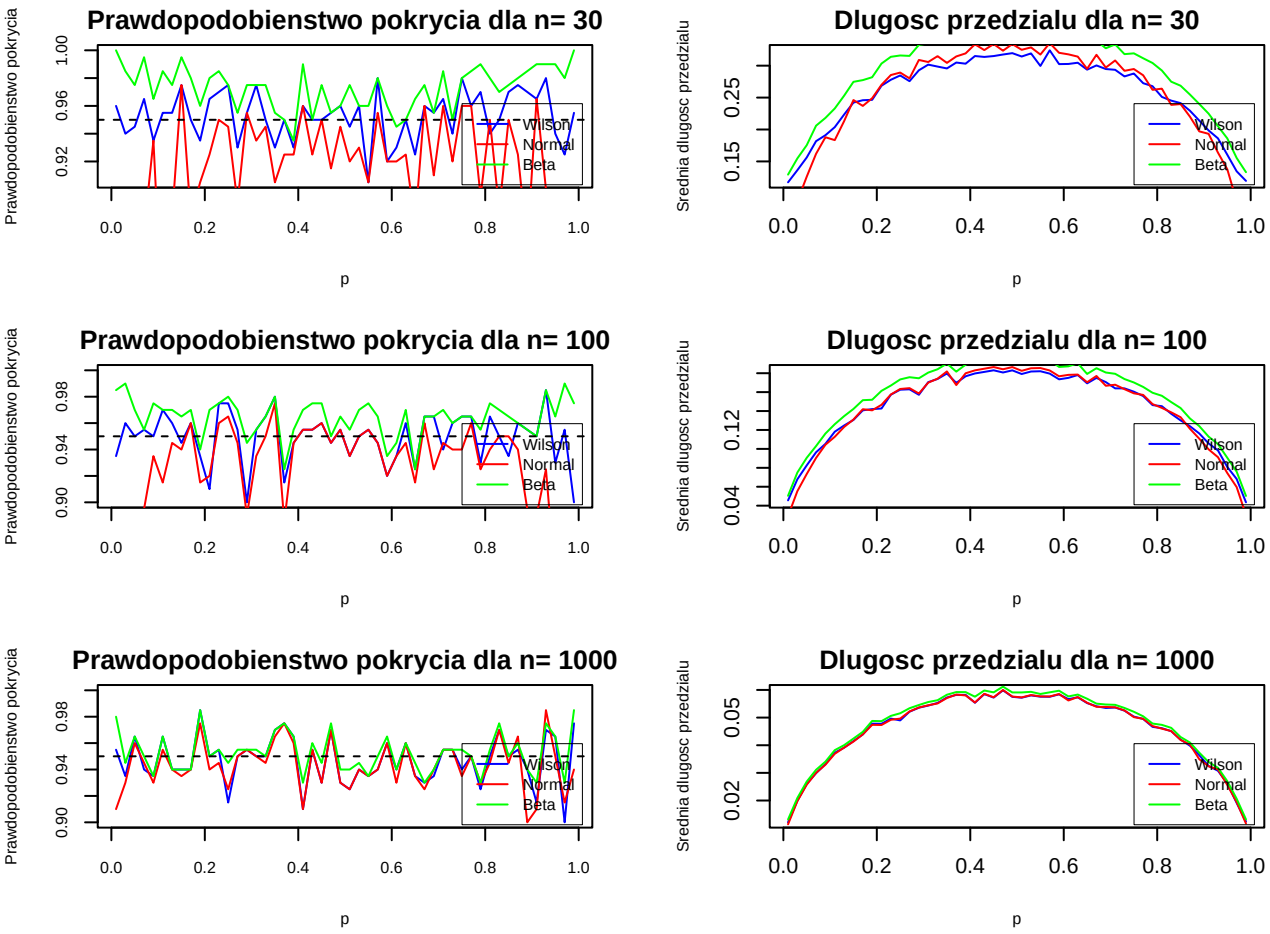
3.2 Zadanie 7

```
## Przedział ufności Cloppera-Pearsona: 0.4896515 0.5093488
```

```
## Przedział ufności Wald'a: 0.4897002 0.5092998
```

```
## Przedział ufności Wilsona: 0.4897023 0.5092981
```

3.3 Zadanie 9



Dla małych próbek (n=30) metoda normalna wykazuje dużą niestabilność, z wyraźnymi spadkami poniżej poziomu pozostałych metod. W miarę wzrostu liczności próby (n=100, n=1000) różnice między metodami się zmniejszają, a wszystkie trzy metody zbliżają się do oczekiwanego poziomu pokrycia. Metoda Wilsona i Beta generalnie utrzymują lepsze prawdopodobieństwo pokrycia, szczególnie dla mniejszych próbek.

Wszystkie metody wykazują podobny kształt – długość przedziału jest największa dla p 0.5p 0.5 i maleje dla wartości bliższych 0 i 1. Dla n=30 metoda normalna daje krótsze przedziały, ale kosztem gorszego prawdopodobieństwa pokrycia. Dla większych próbek długości przedziałów uzyskane różnymi metodami są bardzo podobne.

Przy większych n różnice między metodami stają się mniej istotne – wszystkie trzy metody dają podobne prawdopodobieństwo pokrycia i długości przedziałów. Dla małych próbek metoda Walda wydaje się najmniej stabilna i najgorzej dopasowana.

Metoda Walda nie sprawdza się dobrze dla małych prób – daje niestabilne pokrycie i krótsze przedziały kosztem wiarygodności. Metody Wilsona i Beta są bardziej niezawodne, szczególnie dla mniejszych wartości n.Dla dużych próbek wszystkie metody działają podobnie

4 Część V