Sprawozdanie 1 ADA

Krzysztof Radomski 275968

2 grudnia 2024

Spis treści

1 Zadanie 1

Dane `wage1` pochodzą z pakietu `wooldridge` i są oparte na *Current Population Survey* z 1976 roku. Zostały one zebrane przez Henry'ego Farbera i wykorzystane przez Jeffreya Wooldridge'a, który w latach 1988 pracował z Farberem na MIT.

- Liczba obserwacji: 526.
- Liczba zmiennych: 24.

Opis zmiennych

1. `wage` – średnie wynagrodzenie godzinowe,
2. `educ` – liczba lat edukacji,
3. `exper` – potencjalne doświadczenie zawodowe w latach,
4. `tenure` – staż pracy w aktualnej firmie (w latach),
5. `nonwhite` – zmienna binarna (1 = osoba o innym kolorze skóry niż biała, 0 = biała),
6. `female` – zmienna binarna (1 = kobieta, 0 = mężczyzna),
7. `married` – zmienna binarna (1 = osoba w związku małżeńskim, 0 = inny status),
8. `numdep` – liczba osób na utrzymaniu,
9. `smsa` – zmienna binarna (1 = osoba mieszka w obszarze metropolitalnym, 0 = poza nim),
10. `northcen` – zmienna binarna (1 = mieszkaniec północno-centralnego regionu USA),
11. `south` – zmienna binarna (1 = mieszkaniec południowego regionu USA),
12. `west` – zmienna binarna (1 = mieszkaniec zachodniego regionu USA),
13. `construc` – zmienna binarna (1 = praca w sektorze budowlanym),
14. `ndurman` – zmienna binarna (1 = praca w sektorze produkcji wyrobów nietrwałych),
15. `trcommpu` – zmienna binarna (1 = praca w transporcie, komunikacji lub usługach publicznych),
16. `trade` – zmienna binarna (1 = praca w hurtowniach lub sprzedaży detalicznej),
17. `services` – zmienna binarna (1 = praca w sektorze usługowym),
18. `profserv` – zmienna binarna (1 = praca w sektorze usług profesjonalnych),
19. `profocc` – zmienna binarna (1 = praca w zawodach profesjonalnych),
20. `clerocc` – zmienna binarna (1 = praca w zawodach biurowych),
21. `servocc` – zmienna binarna (1 = praca w zawodach usługowych),
22. `lwage` – logarytm naturalny wynagrodzenia godzinowego,
23. `expersq` – kwadrat zmiennej `exper`,
24. `tenursq` – kwadrat zmiennej `tenure`.

2 Zadanie 2

2.1 (a) Utworzenie zmiennej wageCat na podstawie kwartylów zmiennej wage

```
wage_quartiles <- quantile(wage1$wage, probs = c(0, 0.25, 0.5, 0.75, 1))
wageCat <- cut(wage1$wage, breaks = wage_quartiles, include.lowest = TRUE,
               labels = c("1", "2", "3", "4"))
```

2.2 (b) Utworzenie zmiennych educCat i experCat na podstawie tercyli odpowiednich zmiennych

```
educ_terciles <- quantile(wage1$educ, probs = c(0, 1/3, 2/3, 1))
educCat <- cut(wage1$educ, breaks = educ_terciles, include.lowest = TRUE,
               labels = c("Low", "Medium", "High"))

exper_terciles <- quantile(wage1$exper, probs = c(0, 1/3, 2/3, 1))
experCat <- cut(wage1$exper, breaks = exper_terciles, include.lowest = TRUE,
                labels = c("Low", "Medium", "High"))
```

2.3 (c) Wybór zmiennych binarnych nonwhite, female, married, smsa

```
nonwhite <- wage1$nonwhite
female <- wage1$female
married <- wage1$married
smsa <- wage1$smsa
```

2.4 (d) Utworzenie zmiennej region na podstawie zmiennych northcen, south i west

```
region <- ifelse(wage1$northcen == 1, "North Central",
                 ifelse(wage1$south == 1, "South",
                        ifelse(wage1$west == 1, "West", "Other")))
```

2.5 (e) Połączenie wszystkiego w nową ramkę danych

```
new_data <- data.frame(wageCat, educCat, experCat, nonwhite,
                       female, married, smsa, region)
```

3 Zadanie 3

3.1 (a) Tablice liczności i częstości dla zmiennych region, female, married

Tabela 1: Liczności i czestosci dla zmiennej region

	region	liczność	częstość
1	North Central	132	0.25
2	Other	118	0.22
3	South	187	0.36
4	West	89	0.17

Tabela 2: Liczności i czestosci dla zmiennej married

	region	liczność	częstość
1	0	206	0.39
2	1	320	0.61

Tabela 3: Liczności i czestosci dla zmiennej female

	region	liczność	częstość
1	0	274	0.52
2	1	252	0.48

Dane wskazują, że ludzie biorący udział w badaniu w podobnej ilości mieszkają w każdym regione. Około co trzecia osoba mieszka na południu, a mniej niż jedna piąta na wschodzie. Większość (bo 61%) jest w związku małżeńskim. Możemy też zauważyć, że w badaniu wzięło udział mniej więcej tyle samo mężczyzn, co kobiet.

3.2 (b) Tabele wielodzielcze (liczności i częstości) dla wybranych zmiennych

Tabela 4: Liczności i częstości dla zmiennych wageCat i female

Wage Category	Gender	Frequency	Proportion
1	Female	39	0.29
1	Male	53	0.40
2	Female	79	0.60
2	Male	103	0.79
3	Female	94	0.71
3	Male	78	0.60
4	Female	52	0.40
4	Male	28	0.21

Tabela 5: Liczności i częstości dla zmiennych wageCat i married

Wage Category	Marital Status	Frequency	Proportion
1	Married	79	0.59
1	Not Married	50	0.38
2	Married	45	0.34
2	Not Married	32	0.24
3	Married	54	0.41
3	Not Married	81	0.62
4	Married	86	0.66
4	Not Married	99	0.76

Tabela 6: Liczności i częstości dla zmiennych wageCat i region

Wage Category	Region	Frequency	Proportion
1	North Central	33	0.25
2	North Central	32	0.24
3	North Central	35	0.27
4	North Central	32	0.24
1	Other	27	0.20
2	Other	30	0.23
3	Other	28	0.21
4	Other	33	0.25
1	South	53	0.40
2	South	53	0.40
3	South	44	0.34
4	South	37	0.28
1	West	20	0.15
2	West	16	0.12
3	West	24	0.18
4	West	29	0.22

Tabela 7: Liczności i częstości dla zmiennych wageCat i smsa

Wage Category	SMSA	Frequency	Proportion
1	0	53	0.40
2	0	38	0.29
3	0	35	0.27
4	0	20	0.15
1	1	80	0.60
2	1	93	0.71
3	1	96	0.73
4	1	111	0.85

Tabela 8: Liczności i częstości dla zmiennych wageCat i educCat

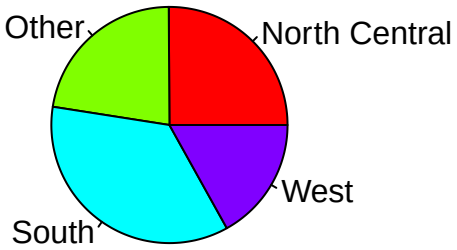
Wage Category	Education Category	Frequency	Proportion
1	Low	111	0.83
2	Low	83	0.63
3	Low	70	0.53
4	Low	50	0.38
1	Medium	8	0.06
2	Medium	12	0.09
3	Medium	10	0.08
4	Medium	9	0.07
1	High	14	0.11
2	High	36	0.27
3	High	51	0.39
4	High	72	0.55

Tabela 9: Liczności i częstości dla zmiennych wageCat i experCat

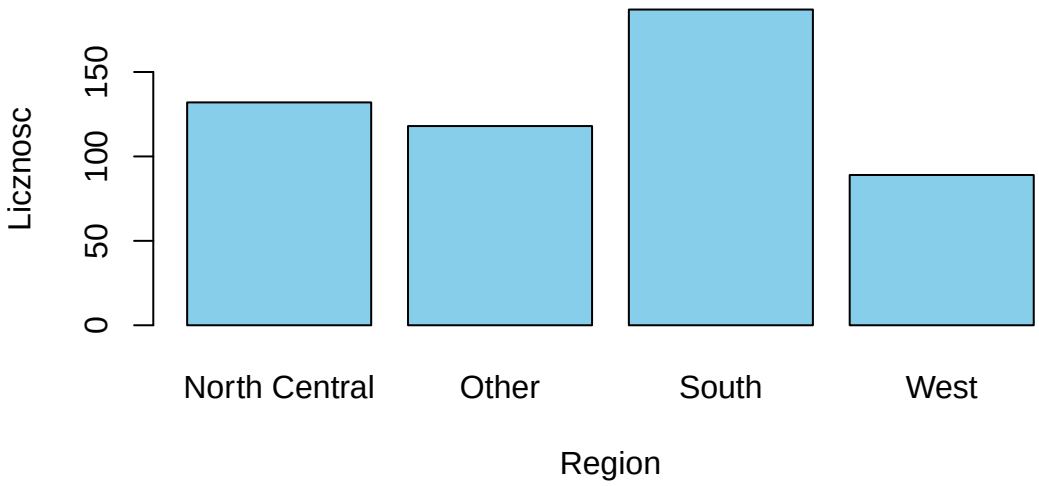
Wage Category	Experience Category	Frequency	Proportion
1	Low	64	0.48
2	Low	55	0.42
3	Low	42	0.32
4	Low	27	0.21
1	Medium	26	0.20
2	Medium	40	0.31
3	Medium	56	0.43
4	Medium	51	0.39
1	High	43	0.32
2	High	36	0.27
3	High	33	0.25
4	High	53	0.40

3.3 (c) Wykresy kołowy i słupkowy dla zmiennej region

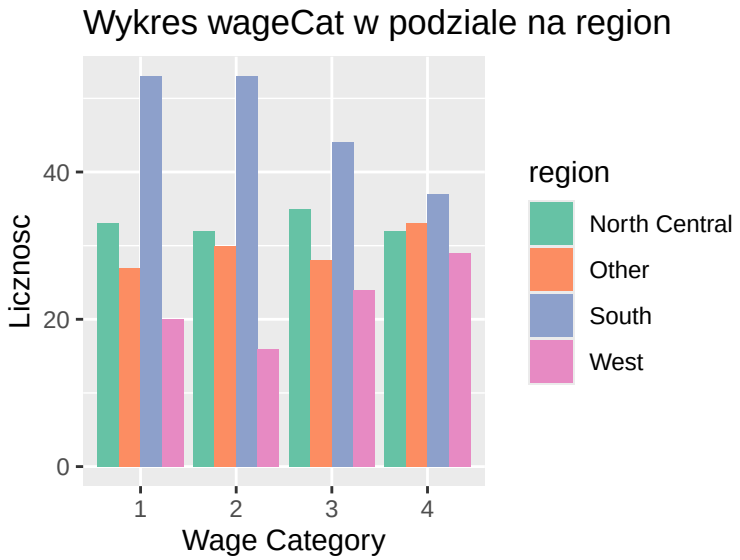
Wykres kolowy dla zmiennej region



Wykres slupkowy dla zmiennej region



3.4 (d) Wykresy wageCat w podziale na region

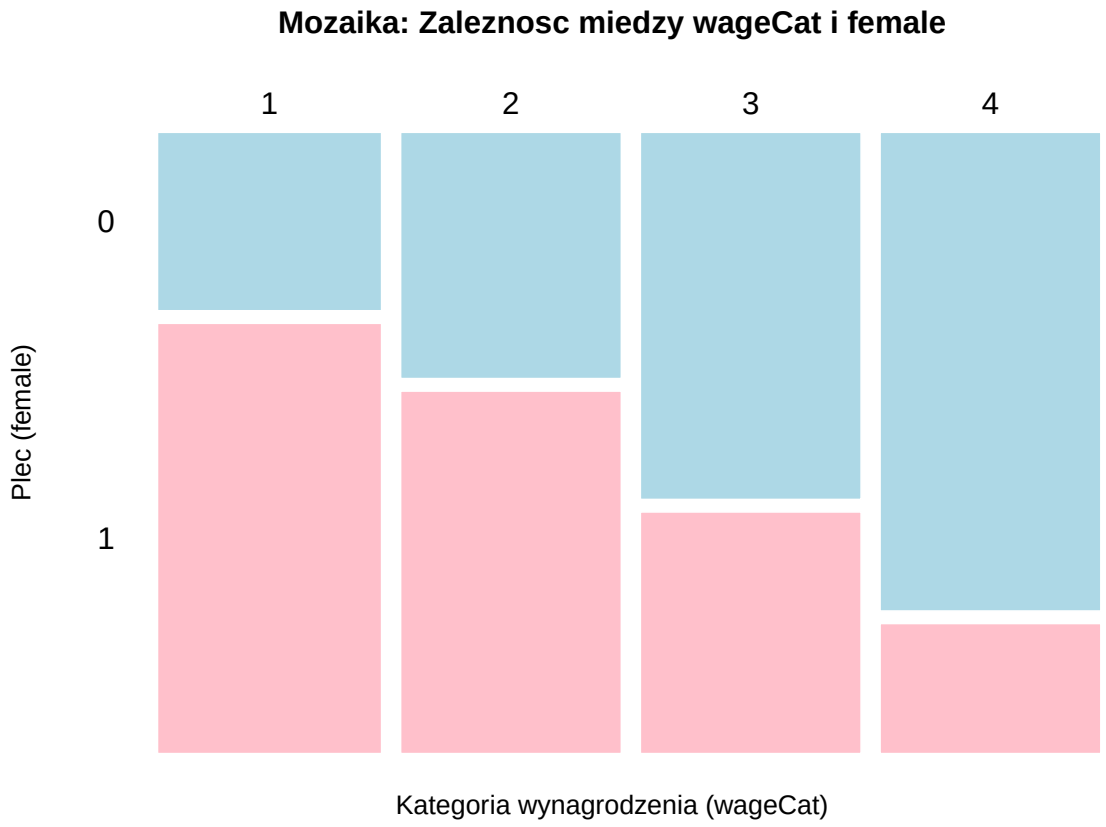


Z wykresu możemy dostrzec niewielką dysproporcje w zarobkach. Widzimy, że ludzie z południa przeciętnie zarabiają mniej od reszty, natomiast mieszkańcy wchodniego regionu, więcej.

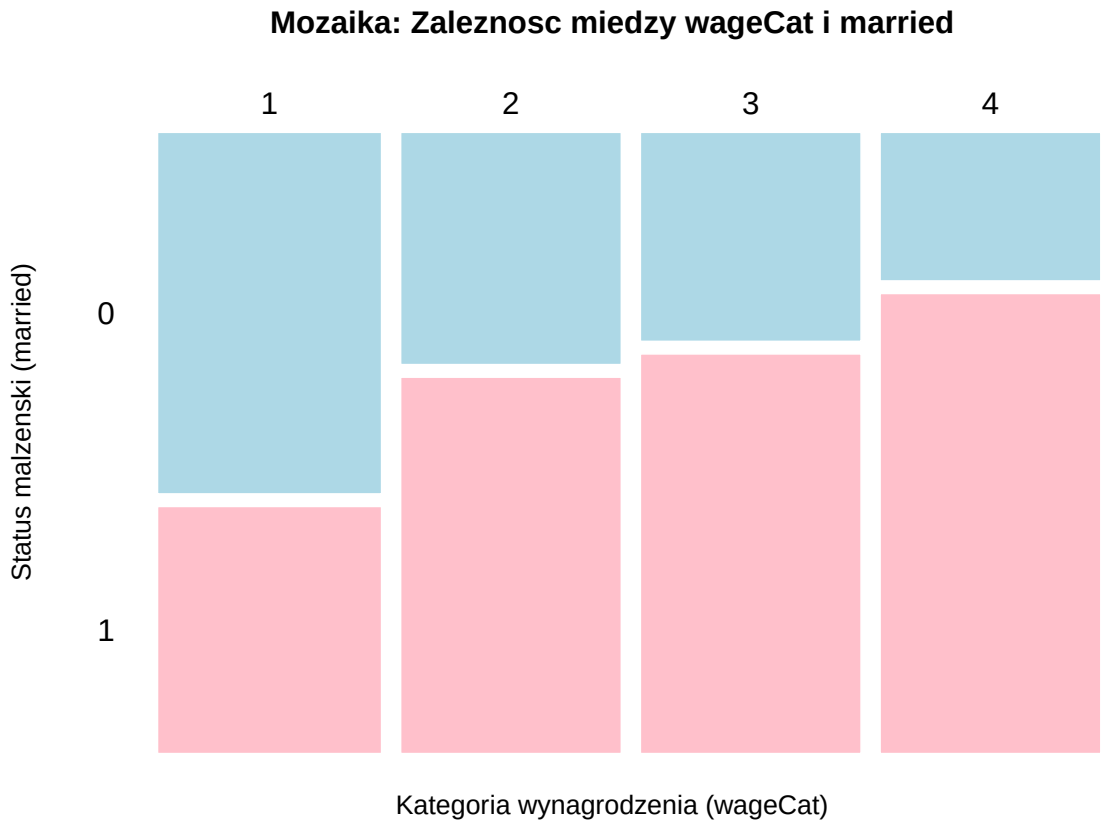
3.5 (e) Wykresy mozaikowe dla par zmiennych z punktu (b)

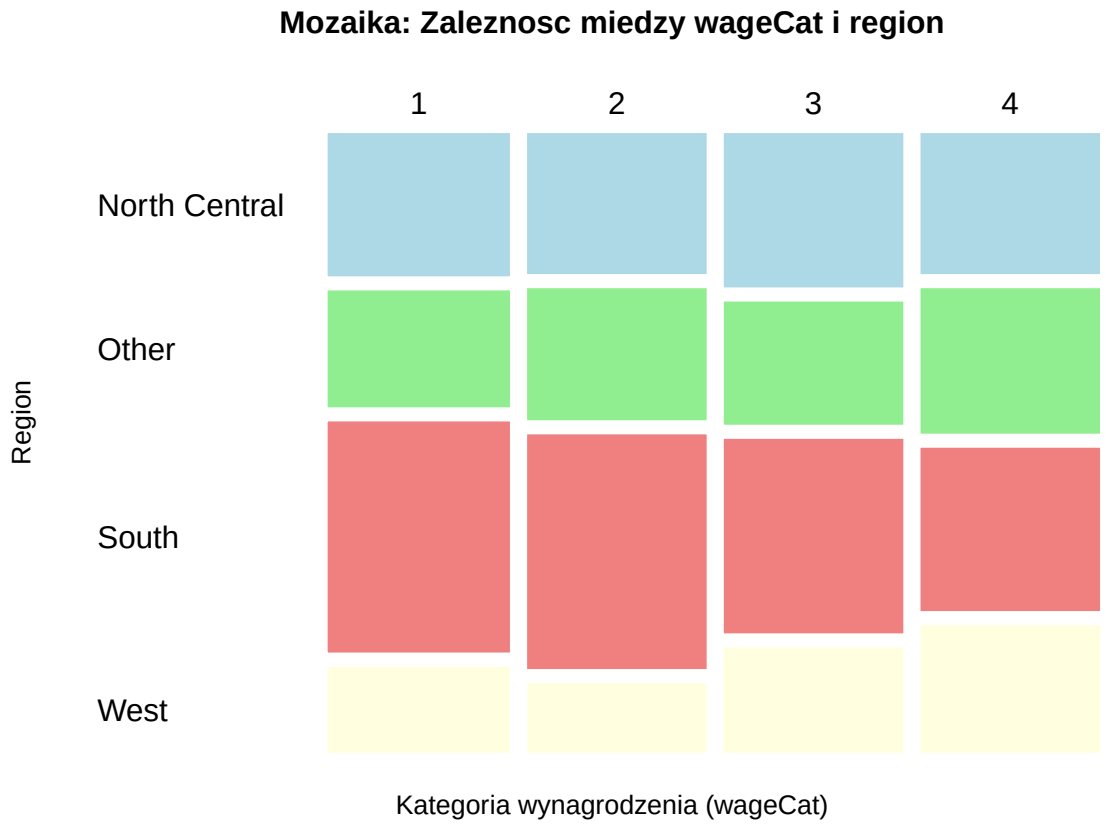
Zmienna WageCat grupuje badanych względem zarobków, gdzie 1 oznacza zbiór ludzi najmniej zamożnych, zaś 4 - najbogatszych. Zaś zmienna female przyjmuje wartość 0 dla mężczyzn, 1 dla kobiet.

Z poniższego wykresu mozaikowego, możemy wyciągnąć wniosek, że im bogatsza grupa, tym mniej znajduje się w niej kobiet względem mężczyzn.

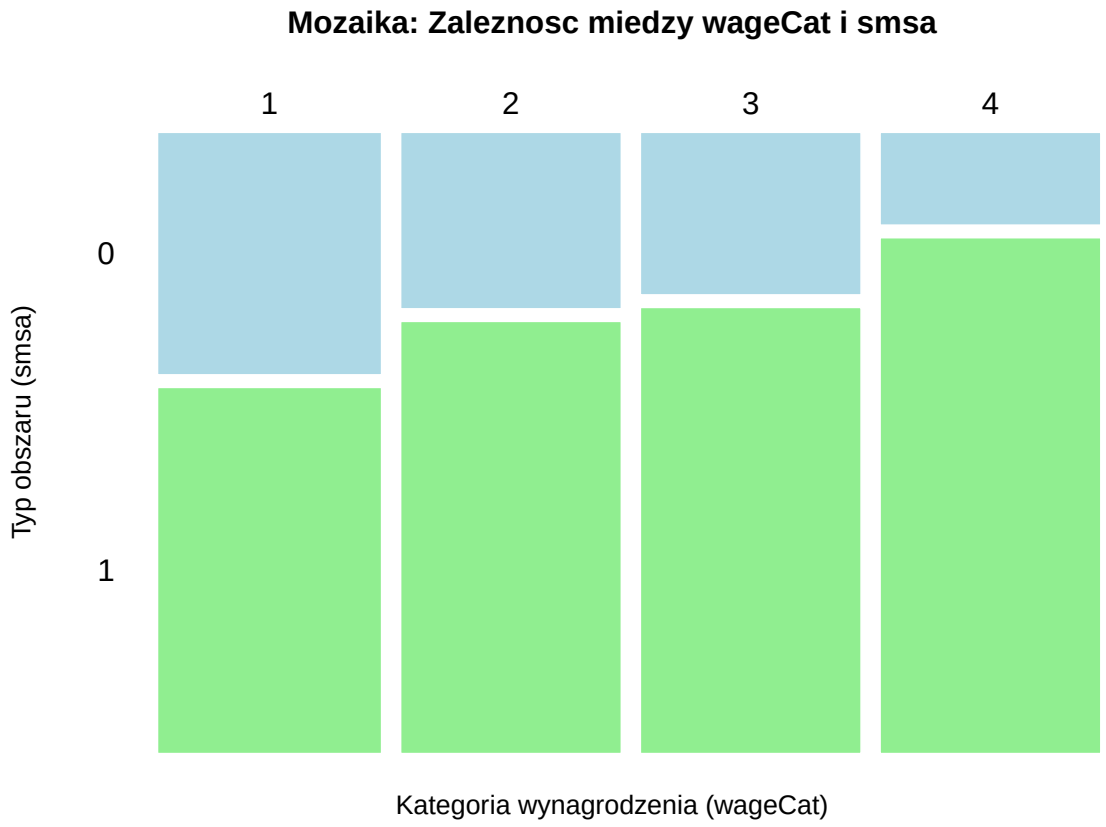


Możemy również wnioskować, że liczba osób w związku małżeńskim, oznaczonych 1 przez zmienną married, rośnie wraz z zarobkami.

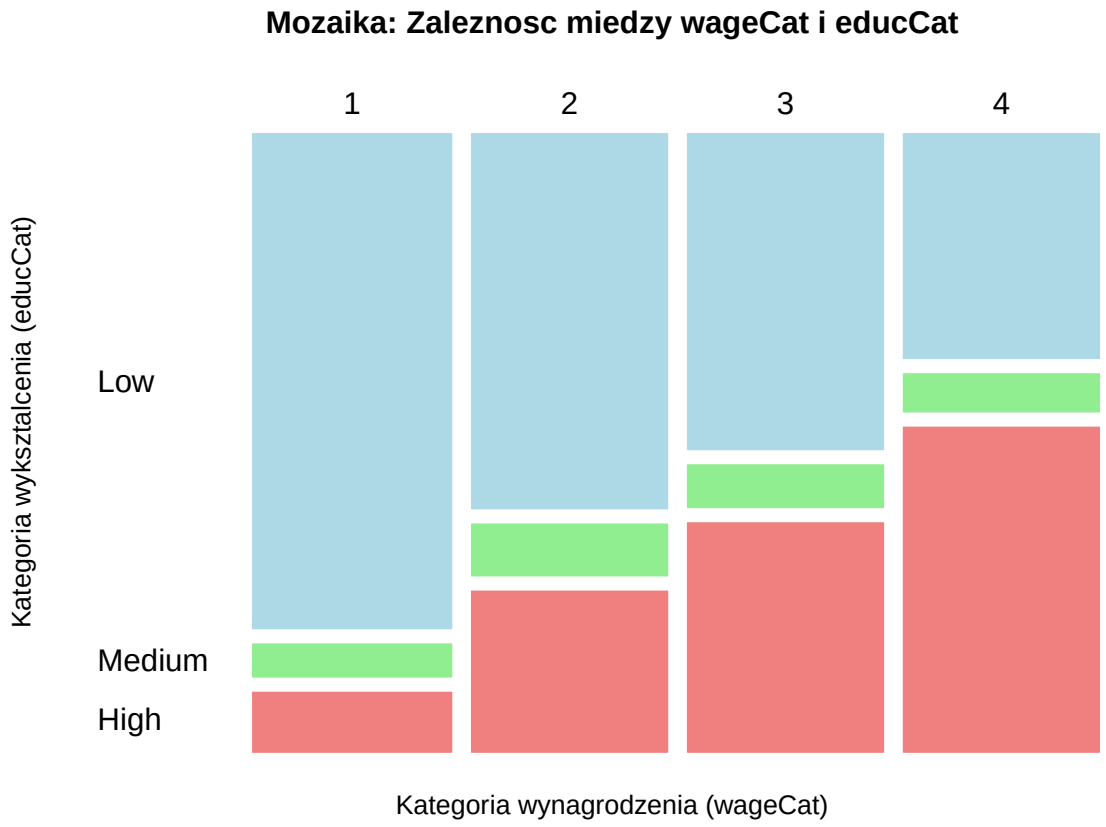

Mozaika: Zaleznosc miedzy wageCat i married
1
2
3
4
0
1
Status malzenski (married)
Kategoria wynagrodzenia (wageCat)

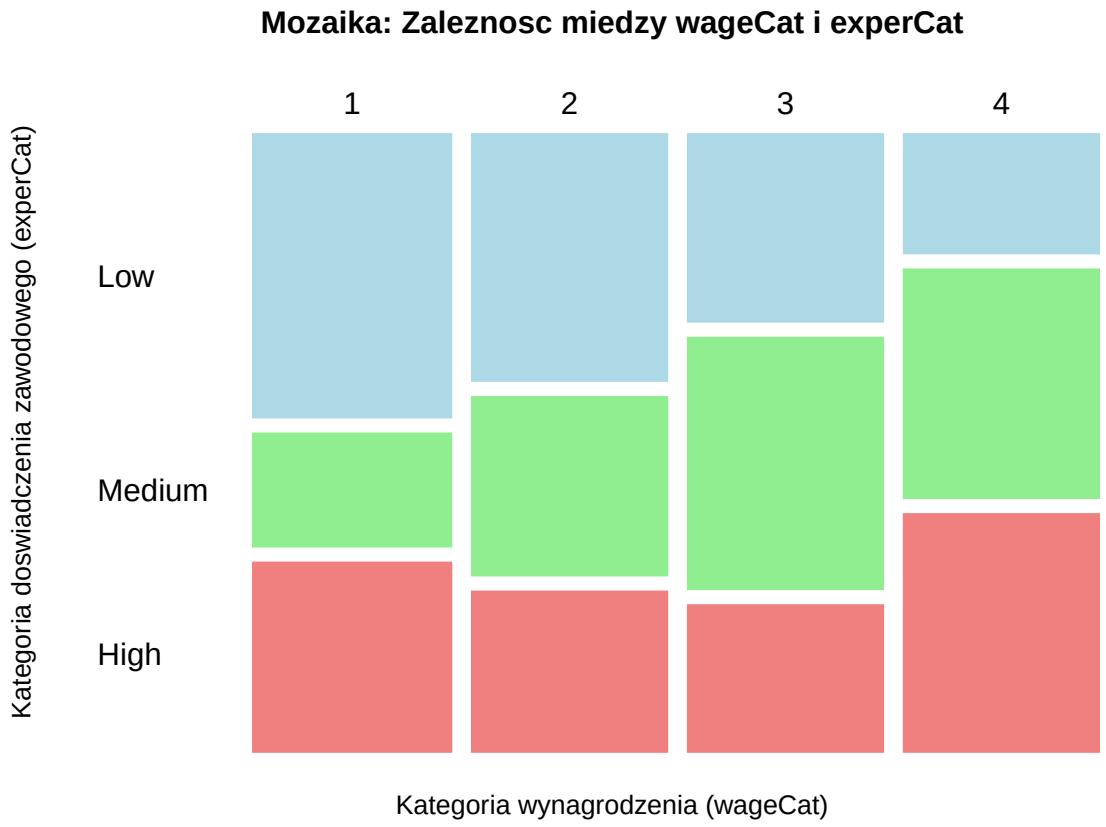


Widać też korelację pomiędzy zmienną wageCat, a zmienną smsa



Widoczna jest również zależność pomiędzy wykształceniem, a zarobkami. Wraz z większym wykształceniem zarobki wzrastają


Mozaika: Zaleznosc miedzy wageCat i educCat
1
2
3
4
Kategoria wykształcenia (educCat)
Low
Medium
High
Kategoria wynagrodzenia (wageCat)



4 Zadanie 4

Opis symulacji

W zadaniu przeprowadzono symulacje przedziałów ufności dla różnych metod estymacji przedziałowej (Clopper-Pearson, Wald i Wilson) dla danych dwumianowych. Symulacja została przeprowadzona dla trzech wartości rozmiaru próby ($n = 30, 50, 100$) i trzech wartości prawdopodobieństwa sukcesu ($p = 0.1, 0.5, 0.9$) przy poziomie ufności wynoszącym 0.95.

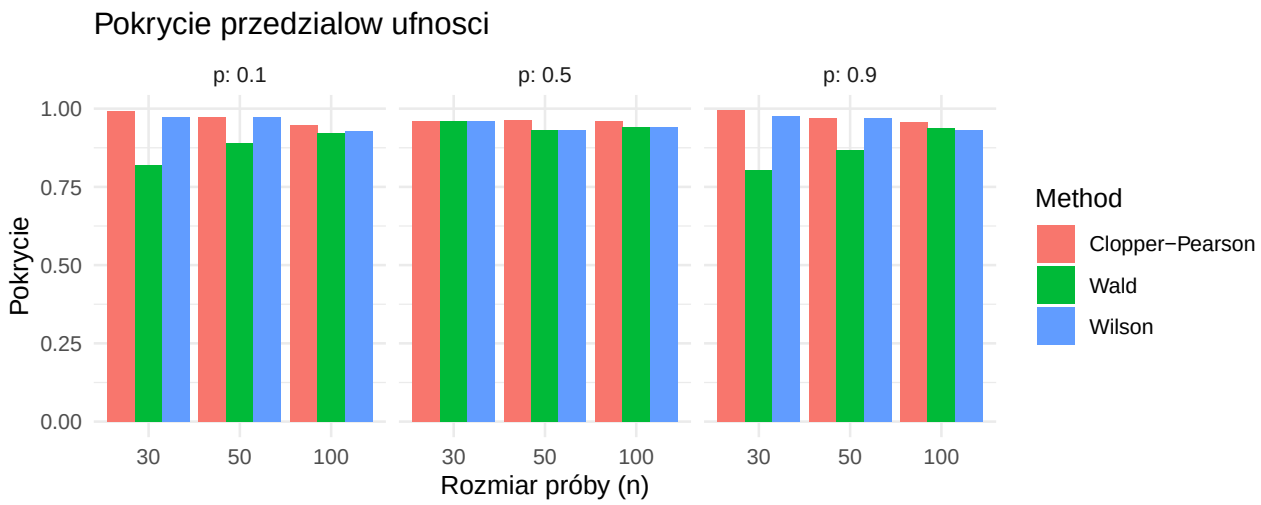
Kod symulacji

```
n_values <- c(30, 50, 100)        # Rozmiary próby
p_values <- c(0.1, 0.5, 0.9)      # Prawdopodobieństwa sukcesu
confidence_level <- 0.95           # Poziom ufności
num_simulations <- 1000

simulate_intervals <- function(n, p, confidence_level, num_simulations) {
  coverage_results <- data.frame(Method = character(), Coverage = numeric(),
                                 Avg_Length = numeric(), stringsAsFactors = FALSE)

  clopper_pearson_coverages <- numeric(num_simulations)
```

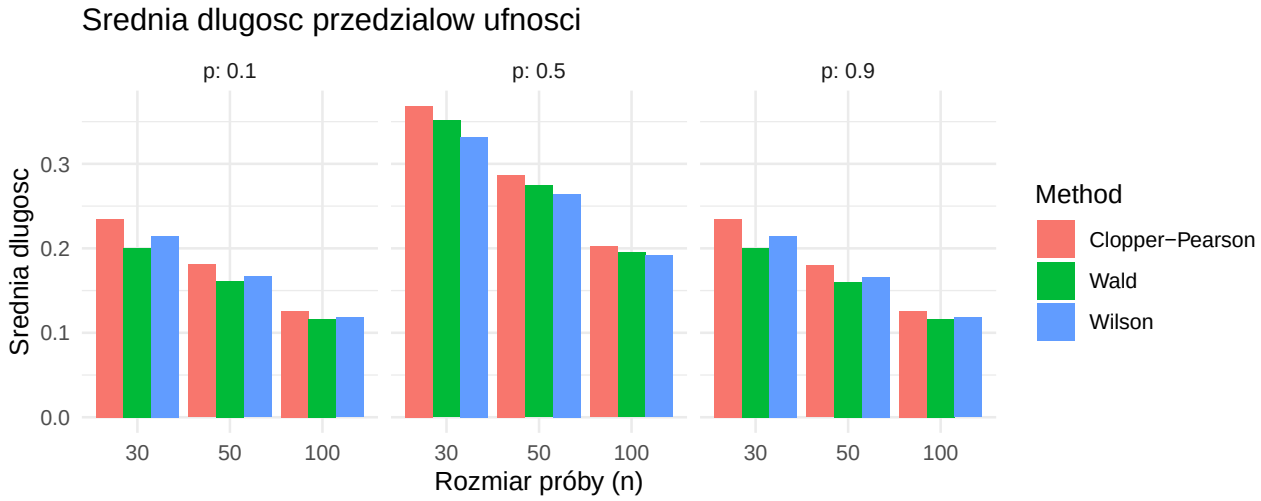
```
  wald_coverages <- numeric(num_simulations)
  wilson_coverages <- numeric(num_simulations)

  clopper_pearson_lengths <- numeric(num_simulations)
  wald_lengths <- numeric(num_simulations)
  wilson_lengths <- numeric(num_simulations)

  for (i in 1:num_simulations) {
    x <- rbinom(1, n, p)

    cp_interval <- binom.confint(x, n, conf.level = confidence_level,
                                 methods = "exact")[, c("lower", "upper")]
    wald_interval <- binom.confint(x, n, conf.level = confidence_level,
                                 methods = "asymptotic")[, c("lower", "upper")]
    wilson_interval <- binom.confint(x, n, conf.level = confidence_level,
                                 methods = "wilson")[, c("lower", "upper")]

    clopper_pearson_coverages[i] <- as.numeric(cp_interval$lower <= p &
                                                 cp_interval$upper >= p)
    wald_coverages[i] <- as.numeric(wald_interval$lower <= p &
                                      wald_interval$upper >= p)
    wilson_coverages[i] <- as.numeric(wilson_interval$lower <= p &
                                        wilson_interval$upper >= p)

    clopper_pearson_lengths[i] <- cp_interval$upper - cp_interval$lower
    wald_lengths[i] <- wald_interval$upper - wald_interval$lower
    wilson_lengths[i] <- wilson_interval$upper - wilson_interval$lower
  }

  coverage_results <- rbind(
    coverage_results,
    data.frame(Method = "Clopper-Pearson",
               Coverage = mean(clopper_pearson_coverages),
               Avg_Length = mean(clopper_pearson_lengths)),
    data.frame(Method = "Wald",
               Coverage = mean(wald_coverages),
               Avg_Length = mean(wald_lengths)),
    data.frame(Method = "Wilson",
               Coverage = mean(wilson_coverages),
               Avg_Length = mean(wilson_lengths))
  )

  return(coverage_results)
}

simulation_results <- data.frame(n = integer(), p = numeric(),
                                 Method = character(),
                                 Coverage = numeric(), Avg_Length = numeric(),
```

```
                                    stringsAsFactors = FALSE)

for (n in n_values) {
  for (p in p_values) {
    results <- simulate_intervals(n, p, confidence_level, num_simulations)
    results$n <- n
    results$p <- p
    simulation_results <- rbind(simulation_results, results)
  }
}
```

Wyniki



Rysunek 1: Pokrycie przedziałów ufności dla różnych metod



Rysunek 2: Średnia długość przedziałów ufności dla różnych metod

Na podstawie wyników symulacji można zauważyć, że niezależnie od wartości parametru **p** oraz wielkości próbki najlepsze pokrycie mają przedziały tworzone metodą Cloppera-Pearsona. Niewiele gorzej radzą sobie przedziały Wilsona, różnica jest mała, jednak widoczna. Przy niewielkich próbkach słabo radzą sobie przedziały Walda, z racji tego, że są asymptotyczne. Istotnym przy wyborze odpowiedniej matody wyznaczania przedziałów ufności jest jego długość. Chcemy, aby przedział był w miare krótki, ale oczywiście z dużym prawdopodobieństwem poprawnie wskazywał szukany, nieznany parametr θ. Niestety przedziały Cloppera-Pearsona są długie w porównaniu do pozostałych, także niezależnie od parametru i rozmiaru próby, najkrótsze zaś są Wilsona i to one, według danych pozyskanych z symulacji są najlepsze, ponieważ są najkrótsze oraz z dobrym prawdopodobieństwem wskazują nieznany parametr.

5 Zadanie 5

5.1 (a)wyznaczyć realizacje przedziałów ufności, na poziomie 0.95, że losowo wybrana osoba z badanej populacji i w podgrupach ze względu na skategoryzowane wynagrodzenie jest kobietą

```r
# Oblicz liczbę kobiet w każdej kategorii wageCat
table_wageCat_female <- table(new_data$wageCat, new_data$female)

# Oblicz proporcję kobiet w każdej kategorii wageCat
prop_wageCat_female <- prop.table(table_wageCat_female, margin = 1)

# Oblicz liczbę kobiet i prób w całej populacji
total_successes <- sum(table_wageCat_female[, 2])  # Liczba kobiet w całej populacji
total_trials <- sum(table_wageCat_female)  # Liczba wszystkich osób w całej populacji

# Funkcja do obliczania przedziałów ufności dla trzech metod
calculate_ci <- function(successes, trials) {
  # Wilson
  ci_wilson <- binom.confint(successes, trials, conf.level = 0.95,
                             methods = "wilson")

  # Clopper-Pearson
  ci_clopper <- binom.confint(successes, trials, conf.level = 0.95,
                              methods = "exact")

  # Wald
  ci_wald <- binom.confint(successes, trials, conf.level = 0.95,
                           methods = "asymp")

  # Połącz wyniki w jedną tabelę
  result <- data.frame(
    Wilson_Lower = ci_wilson$lower,
    Wilson_Upper = ci_wilson$upper,
    Clopper_Pearson_Lower = ci_clopper$lower,
```

```
    Clopper_Pearson_Upper = ci_clopper$upper,
    Wald_Lower = ci_wald$lower,
    Wald_Upper = ci_wald$upper
  )

  return(result)
}

# Oblicz przedziały ufności dla każdej kategorii wageCat
ci_results <- lapply(1:nrow(prop_wageCat_female), function(i) {
  successes <- table_wageCat_female[i, 2]  # Liczba kobiet
  trials <- sum(table_wageCat_female[i, ]) # Liczba osób w danej kategorii wageCat

  # Oblicz przedziały ufności
  calculate_ci(successes, trials)
})

# Zmień listę wyników na ramkę danych dla łatwiejszej analizy
ci_df <- do.call(rbind, ci_results)
rownames(ci_df) <- rownames(prop_wageCat_female)

# Oblicz przedziały ufności dla całej populacji
total_ci <- calculate_ci(total_successes, total_trials)

# Dodaj wiersz dla całej populacji na końcu
ci_df <- rbind(ci_df, total_ci)
rownames(ci_df)[nrow(ci_df)] <- "Total Population"
```

Tabela 10: Przedziały ufności dla prawdopodobieństwa, że losowo wybrana osoba jest kobietą w zależności od kategorii wynagrodzenia (Wilson)

	Wilson_Lower	Wilson_Upper
1	0.6245	0.7775
2	0.5098	0.6756
3	0.3172	0.4825
4	0.1522	0.2916
Total Population	0.4367	0.5218

Tabela 11: Przedziały ufności dla prawdopodobieństwa, że losowo wybrana osoba jest kobietą w zależności od kategorii wynagrodzenia (Clopper-Pearson)

	Clopper_Pearson_Lower	Clopper_Pearson_Upper
1	0.6216	0.7825
2	0.5062	0.6802
3	0.3126	0.4861
4	0.1470	0.2939
Total Population	0.4357	0.5227

Tabela 12: Przedziały ufności dla prawdopodobieństwa, że losowo wybrana osoba jest kobietą w zależności od kategorii wynagrodzenia (Wald)

	Wald_Lower	Wald_Upper
1	0.6294	0.7841
2	0.5114	0.6795
3	0.3132	0.4807
4	0.1435	0.2839
Total Population	0.4364	0.5218

Porównując długości przedziałów oraz korzystając z wiedzy z poprzedniego zadania, możemy wnioskiować, że najlepszymi wyznaczonymi przedziałami ufności będą te, wyznaczone metodą Wilsona. Zatem odczytując tabelkę prawdopodobienstwo tego, że losowo wybrana osoba z grupy najgorzej zarabiającej na 95% jest warością z przedziału 0.6245 - 0.775. Jeśli chodzi o grupę zarabiającą niewiele więcej, to prawdopodobieństwo jest w przedziale 0.5098 - 0.6756, dla 3-ciej: 0.3172 - 0.4825, dla najbogatszej zaś to prawdopodobieństwo wynosi pomiędzy 0.1522 - 0.2916.

5.2 (b) Wyznaczyć realizacje przedziałów ufności, na poziomie 0.95, że losowo wybrana osoba z badanej populacji i w podgrupach ze względu na region jest w związku małżeńskim

```r
# Oblicz liczbę osób w związku małżeńskim w każdej kategorii region
table_region_married <- table(new_data$region, new_data$married)

# Oblicz proporcję osób w związku małżeńskim w każdej kategorii region
prop_region_married <- prop.table(table_region_married, margin = 1)

# Oblicz przedziały ufności dla każdej proporcji w podziale na region oraz dla całej p
ci_results <- lapply(1:nrow(prop_region_married), function(i) {
  successes <- table_region_married[i, 2]  # Liczba osób w związku małżeńskim
  trials <- sum(table_region_married[i, ]) # Liczba osób w danej kategorii region

  # Oblicz przedziały ufności przy poziomie ufności 0.95 metodami Wilsona, Cloppera-Pe
  ci_wilson <- binom.confint(successes, trials, conf.level = 0.95,
                             methods = "wilson")
  ci_cp <- binom.confint(successes, trials, conf.level = 0.95,
                         methods = "exact")
  ci_wald <- binom.confint(successes, trials, conf.level = 0.95,
                           methods = "asymptotic")

  # Zwróć wyniki w jednym wierszu z trzema metodami
  return(data.frame(
    Region = rownames(table_region_married)[i],
    Wilson_Lower = ci_wilson$lower, Wilson_Upper = ci_wilson$upper,
    Clopper_Pearson_Lower = ci_cp$lower, Clopper_Pearson_Upper = ci_cp$upper,
    Wald_Lower = ci_wald$lower, Wald_Upper = ci_wald$upper
  ))
```

```
})

# Oblicz dla całej populacji (bez podziału na region)
total_successes <- sum(new_data$married == 1)  # Łączna liczba osób w związku małżeński
total_trials <- nrow(new_data)                 # Całkowita liczba osób w badaniu

ci_total_wilson <- binom.confint(total_successes, total_trials,
                                 conf.level = 0.95, methods = "wilson")
ci_total_cp <- binom.confint(total_successes, total_trials,
                             conf.level = 0.95, methods = "exact")
ci_total_wald <- binom.confint(total_successes, total_trials,
                               conf.level = 0.95, methods = "asymptotic")

# Dodaj wyniki dla całej populacji do wyników dla regionów
ci_results[[nrow(prop_region_married) + 1]] <- data.frame(
  Region = "Total Population",
  Wilson_Lower = ci_total_wilson$lower,
  Wilson_Upper = ci_total_wilson$upper,
  Clopper_Pearson_Lower = ci_total_cp$lower,
  Clopper_Pearson_Upper = ci_total_cp$upper,
  Wald_Lower = ci_total_wald$lower, Wald_Upper = ci_total_wald$upper
)

# Zmień listę wyników na ramkę danych dla łatwiejszej analizy i wyświetlenia
ci_df <- do.call(rbind, ci_results)
```

Tabela 13: Przedziały ufności metodą Wilsona dla związku małżeńskiego w podziale na region

	Region	Wilson Lower	Wilson Upper
1	North Central	0.51	0.68
2	Other	0.47	0.65
3	South	0.59	0.72
4	West	0.48	0.68
5	Total Population	0.57	0.65

Tabela 14: Przedziały ufności metodą Cloppera-Pearsona dla związku małżeńskiego w podziale na region

	Region	Clopper-Pearson Lower	Clopper-Pearson Upper
1	North Central	0.51	0.68
2	Other	0.46	0.65
3	South	0.58	0.73
4	West	0.47	0.69
5	Total Population	0.57	0.65

Tabela 15: Przedziały ufności metodą Walda dla związku małżeńskiego w podziale na region

	Region	Wald Lower	Wald Upper
1	North Central	0.51	0.68
2	Other	0.47	0.65
3	South	0.59	0.73
4	West	0.48	0.69
5	Total Population	0.57	0.65

W tym przypadku przedziały Wilsona, również są najlepszym wyborem, ponieważ są one najkrótsze oraz najtrafniej wskazują rzeczywiste prawdopodbieństwa. Następujące przedziały dla kolejnych regionów to:

1. `North Central` – [0.51 - 0.68]
2. `Other` – [0.47 - 0.65]
3. `South` – [0.59 - 0.72]
4. `West` – [0.48 - 0.68]

6 Zadanie 6

6.1 a) Hipoteza: prawdopodobieństwo, że losowo wybrana osoba z badanej populacji jest w związku małżeńskim, jest większe bądź równe 0.75

```
successes_married <- sum(new_data$married)
trials <- nrow(new_data)

test_a <- binom.test(successes_married, trials, p = 0.75, alternative = "less")
test_a$p.value

## [1] 6.188654e-13
```

Wartość p (p-value) dla hipotezy (a): $6.1886539 \times 10^{-13}$, zatem na poziomie istotności, $\alpha = 0.05$ odrzucamy hipotezę zerową, to znaczy, że prawdopodobieństwo musi być mniejsze niż 0.75.

6.2 b) Hipoteza: prawdopodobieństwo, że losowo wybrana osoba z grupy osób o najwyższym wynagrodzeniu jest kobietą, jest równe 1/2

```
highest_wage_group <- new_data[new_data$wageCat == 4, ]

successes_highest_wage <- sum(highest_wage_group$female)
trials_highest_wage <- nrow(highest_wage_group)
```

```r
test_b <- binom.test(successes_highest_wage, trials_highest_wage, p = 0.5
                     , alternative = "two.sided")
test_b$p.value

## [1] 2.80458e-11
```

Wartość p (p-value) dla hipotezy (b): $2.8045803 \times 10^{-11}$, zatem na poziomie istotności, $\alpha = 0.05$ odrzucamy hipotezę zerową, to znaczy, że prawdopodobieństwo musi być różne od 0.5

6.3 c) Hipoteza: prawdopodobieństwo, że losowo wybrana osoba z grupy osób o najniższym wynagrodzeniu jest kobietą, jest równe 1/2

```r
lowest_wage_group <- new_data[new_data$wageCat == 1, ]

successes_lowest_wage <- sum(lowest_wage_group$female)
trials_lowest_wage <- nrow(lowest_wage_group)

test_c <- binom.test(successes_lowest_wage, trials_lowest_wage, p = 0.5,
                     alternative = "two.sided")
test_c$p.value

## [1] 2.0566e-06
```

Wartość p (p-value) dla hipotezy (c): 2.0566001×10^{-6}, zatem na poziomie istotności, $\alpha = 0.05$ odrzucamy hipotezę zerową, to znaczy, że prawdopodobieństwo musi być różne od 0.5

7 Zadanie 7

7.1 Hipoteza (a): Porównanie wynagrodzeń na najniższym poziomie między regionami

```r
northcen_lowest_wage <- new_data[new_data$region == 'North Central' &
                                   new_data$wageCat == 1, ]
northcen_successes <- nrow(northcen_lowest_wage)
northcen_trials <- sum(new_data$region == 'North Central')

print(paste("North Central Successes:", northcen_successes))

## [1] "North Central Successes: 33"

print(paste("North Central Trials:", northcen_trials))

## [1] "North Central Trials: 132"
```

```r
south_lowest_wage <- new_data[new_data$region == 'South' &
                                new_data$wageCat == 1, ]
south_successes <- nrow(south_lowest_wage)
south_trials <- sum(new_data$region == 'South')

print(paste("South Successes:", south_successes))

## [1] "South Successes: 53"

print(paste("South Trials:", south_trials))

## [1] "South Trials: 187"

test_a <- prop.test(c(northcen_successes, south_successes),
                    c(northcen_trials, south_trials))
print(test_a$p.value)

## [1] 0.5930291
```

Na podstawie wyników testu mogę stwierdzić, że postawiona przeze mnie hipoteza zerowa (czyli, prawdopodobieństwo, że losowo wybrana osoba z regionu północno-centralnego ma wynagrodzenie na najniższym poziomie jest równe prawdopodobieństwu, że losowo wybrana osoba z regionu południowego ma wynagrodzenie na najniższym poziomie) nie ma podstaw do jej odrzucenia na poziomie $\alpha = 0.05$, ponieważ p-value wyniosło w przybliżeniu 0.59

7.2 Hipoteza (b): Porównanie prawdopodobieństwa bycia kobietą w grupach o najniższym i najwyższym wynagrodzeniu

```r
highest_wage_female <- sum(new_data$female[new_data$wageCat == 4])
highest_wage_trials <- sum(new_data$wageCat == 4)

print(paste("Highest Wage Female Count:", highest_wage_female))

## [1] "Highest Wage Female Count: 28"

print(paste("Highest Wage Trials:", highest_wage_trials))

## [1] "Highest Wage Trials: 131"

lowest_wage_female <- sum(new_data$female[new_data$wageCat == 1])
lowest_wage_trials <- sum(new_data$wageCat == 1)

print(paste("Lowest Wage Female Count:", lowest_wage_female))

## [1] "Lowest Wage Female Count: 94"

print(paste("Lowest Wage Trials:", lowest_wage_trials))

## [1] "Lowest Wage Trials: 133"
```

```r
test_b <- prop.test(c(highest_wage_female, lowest_wage_female),
                    c(highest_wage_trials, lowest_wage_trials))
print(test_b$p.value)

## [1] 2.571027e-15
```

W tym teście p wartość wyniosła 2.571027×10^{-15}, zatem na poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę zerową. Oznacza to, że prawdopodobieństwo że losowo wybrana osoba z grupy osób o najwyższym wynagrodzeniu jest kobietą NIE jest równe prawdopodobieństwu, że losowo wybrana osoba o najniższym wynagrodzeniu jest kobietą

8 Zadanie 8

```r
asymptotic_test<-function(x,n){
  z<-(x[1]/n[1]-x[2]/n[2])/sqrt((1/n[1]+1/n[2])*(x[1]+x[2])/
                                  (n[1]+n[2])*(1-(x[1]+x[2])/(n[1]+n[2])))
  p.value<-2*(1-pnorm(abs(z)))
  return(p.value)}
```

Zakładamy, że:

$$X_1 \sim \mathrm{B}(p_1, n_1) \quad \text{i} \quad X_2 \sim \mathrm{B}(p_2, n_2),$$

gdzie n_1 i n_2 to liczności obu prób.

Hipotezy są następujące:

$$H_0 : p_1 = p_2, \quad H_1 : p_1 \neq p_2.$$

Statystyka testowa

Funkcja najpierw oblicza estymatory częstości sukcesu:

$$\hat{p}_1 = \frac{\sum x_1}{n_1}, \quad \hat{p}_2 = \frac{\sum x_2}{n_2}.$$

Następnie wyznacza połączony estymator prawdopodobieństwa sukcesu:

$$\hat{p} = \frac{\sum x_1 + \sum x_2}{n_1 + n_2}.$$

Statystyka testowa Z jest obliczana według wzoru:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Poziom krytyczny

Zakładając, że Z ma rozkład asymptotyczny normalny, wartość p-value jest obliczana jako:

$$\text{p-value} = 2 \cdot (1 - \Phi(|Z|)),$$

gdzie Φ to dystrybuanta rozkładu normalnego standardowego.

Wynik testu

Funkcja zwraca wartość statystyki Z, poziom krytyczny (p-value) oraz wynik testu, wskazujący, czy hipoteza zerowa powinna zostać odrzucona dla przyjętego poziomu istotności α.

8.1 Hipoteza (a): Porównanie wynagrodzeń na najniższym poziomie między regionami

```
north_central_data <- new_data[new_data$region == "North Central" &
                                 new_data$wageCat == 1, ]
south_data <- new_data[new_data$region == "South" & new_data$wageCat == 1, ]

x_1 <- c(nrow(north_central_data), nrow(south_data))
n_1 <- c(nrow(new_data[new_data$region == "North Central", ]),
         nrow(new_data[new_data$region == "South", ]))

p_value_asymptotic_1 <- asymptotic_test(x_1, n_1)

p_value_prop_1 <- prop.test(x_1, n_1)$p.value


results_a <- list(Hipoteza_1 = list(Asymptotyczny_test = p_value_asymptotic_1,
                                    Prop_test = p_value_prop_1))
print(results_a)

## $Hipoteza_1
## $Hipoteza_1$Asymptotyczny_test
## [1] 0.507623
##
## $Hipoteza_1$Prop_test
## [1] 0.5930291
```

8.2 Hipoteza (b): Porównanie prawdopodobieństwa bycia kobietą w grupach o najniższym i najwyższym wynagrodzeniu

```
highest_wage_data <- new_data[new_data$wageCat == 4, ]
lowest_wage_data <- new_data[new_data$wageCat == 1, ]

x_2 <- c(sum(highest_wage_data$female == 1), sum(lowest_wage_data$female == 1))
n_2 <- c(nrow(highest_wage_data), nrow(lowest_wage_data))

p_value_asymptotic_2 <- asymptotic_test(x_2, n_2)

p_value_prop_2 <- prop.test(x_2, n_2)$p.value
```

```r
results_b <- list(Hipoteza_2 = list(Asymptotyczny_test = p_value_asymptotic_2,
                                    Prop_test = p_value_prop_2))
print(results_b)

## $Hipoteza_2
## $Hipoteza_2$Asymptotyczny_test
## [1] 8.881784e-16
##
## $Hipoteza_2$Prop_test
## [1] 2.571027e-15
```

Jak widać, w obu przypadkach p wartość liczona 'naszą' funckją nie różni się zbyt wiele od tej z pakietu R. W podpunkcie *a*) 'nasza' p wartość wyniosła 0.507623, zaś ta, którą wyliczyła wbudowana funkcja 0.5930291, zatem w obu przypadkach zdecydowalibyśmy, że nie ma podstaw do odrzucenia H_0. W podpunkcie *b*) wyszło 8.881784×10^{-16}, a wartość ta niewiele się różni od tej z wbudowanej funkcji, która wyniosła 2.571027×10^{-15}. W tym przypadku również byśmy zdecydowali o odrzuceniu hipotezy zaerowej, zatem stworzona przez nas funkcja sprawdziła się w tym przypadku dobrze.