PSP – Raport 5

Nela Tomaszewicz

Czerwiec 2020

1 Zadanie 1

Korzystamy ze zbioru `income.txt` zawierającego dane dotyczące 55 899 pracowników z USA.

1.1 Przybliżony przedział ufności Agrestiego-Coulla

Losujemy 200-elementową próbę ze zbioru `income.txt`, a następnie konstruujemy 95% przedzial ufności metodą Agrestiego-Coulla dla frakcji osób z wykształceniem wyższym (p_W), frakcji kobiet (p_K) oraz frakcji osób zatrudnionych w sektorze prywatnym (p_P). Przybliżony przedział ufności Agrestiego-Coulla konstruujemy, zaczynając od wyznaczenia środka przedziału, czyli zmodyfikowanej frakcji:

$$\tilde{p} = \frac{Y + 0.5(Z_{\frac{\alpha}{2}})^2}{n + (Z_{\frac{\alpha}{2}})^2}, \tag{1}$$

gdzie Y stanowi liczbę sukcesów w próbie, czyli np. liczbę osób będących zatrudnionych w sektorze prywatnym dla zmiennej p_W Następnie konstruujemy SE według wzoru:

$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + (Z_{\frac{\alpha}{2}})^2}} \tag{2}$$

Na końcu otrzymujemy przedział ufności na poziomie istotności α:

$$\tilde{p} \pm Z_{\frac{\alpha}{2}} SE_{\tilde{p}} \tag{3}$$

Zadanie realizuję następującą funkcją:

```
library(dplyr)
n <- 200
alpha <- 0.05
income_random <- sample_n(income, n)
z <- qnorm(1 - alpha/2)

# zliczamy sukcesy
Y_pw <- sum(income_random$wykszt >= 5)

Y_pk <- sum(income_random$plec == 2)
```

```r
Y_pp <- sum(income_random$sektor == 5)

agresti_coull <- function(Y) {
  p_tilde <- (Y + 0.5*z^2)/(n + z^2)
  SE <- sqrt(p_tilde*(1 - p_tilde)/(n + z^2))
  left <- p_tilde - z*SE
  right <- p_tilde + z*SE

  cat("Przedzial ufnosci: [", left, ",", right, "]")
}
```

Przedział ufności dla frakcji osób z wykształceniem wyższym (p_W):

```
## Przedzial ufnosci: [ 0.2360228 , 0.3617038 ]
```

Rzeczywista wartość (wyznaczona dla całego zbioru danych) wynosi około 0.297 i zawiera się w przedziale.

Przedział ufności dla frakcji kobiet (p_K):

```
## Przedzial ufnosci: [ 0.3681464 , 0.5043035 ]
```

Rzeczywista wartość frakcji kobiet wynosi około 0.477 i zawiera się w przedziale.

Przedział ufności dla frakcji osób zatrudnionych w sektorze prywatnym (p_P):

```
## Przedzial ufnosci: [ 0.6591732 , 0.7823464 ]
```

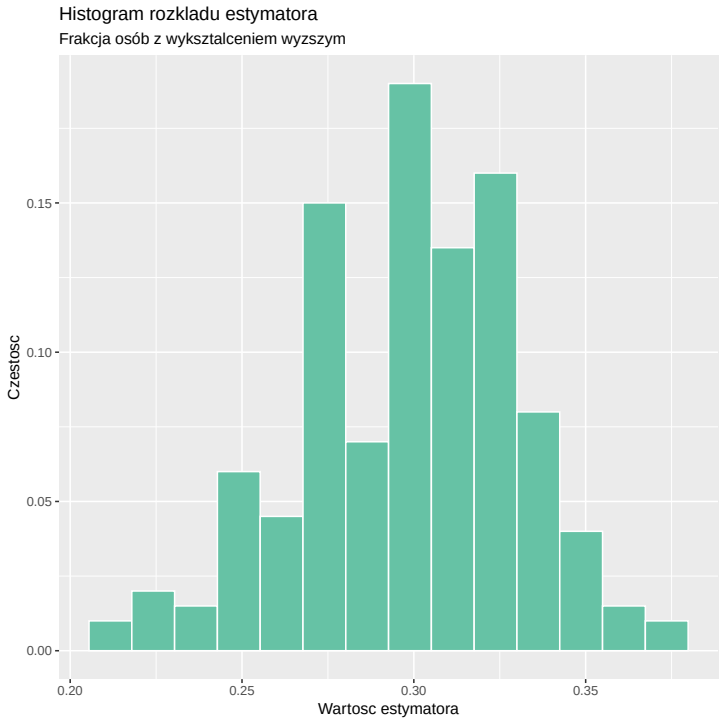
Rzeczywista wartość frakcji osób zatrudnionych w sektorze prywatnym wynosi około 0.732 i również zawiera się w przedziale.

1.2 Powtórzenie 200 razy

Powtarzamy wyznaczanie przedziałów ufności metodą Agrestiego-Coulla 200 razy, a następnie konstruujemy histogramy rozkładów estymatorów $\tilde{p_W}$, $\tilde{p_K}$, $\tilde{p_P}$ i wyznaczamy częstość pokrycia rzeczywistej wartości frakcji przez przedział ufności oraz szerokość przedziałów.

Dla frakcji p_W:

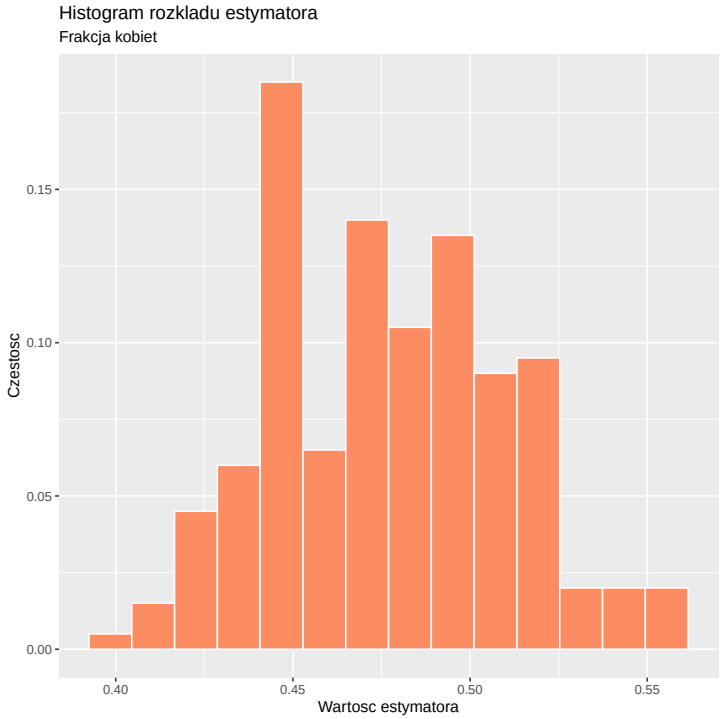
```
## Srednia szerokosc przedzialu:  0.125302
## Czestosc zawierania:  0.95
```



Najczęściej występującą wartością estymatora dla p_W jest około 0.3, co jest poprawne, ponieważ wartość rzeczywista wynosi właśnie około 0.3. Częstość zawierania wynosi 0.95, czyli dokładnie tyle ile poziom ufności.

Dla frakcji p_K:

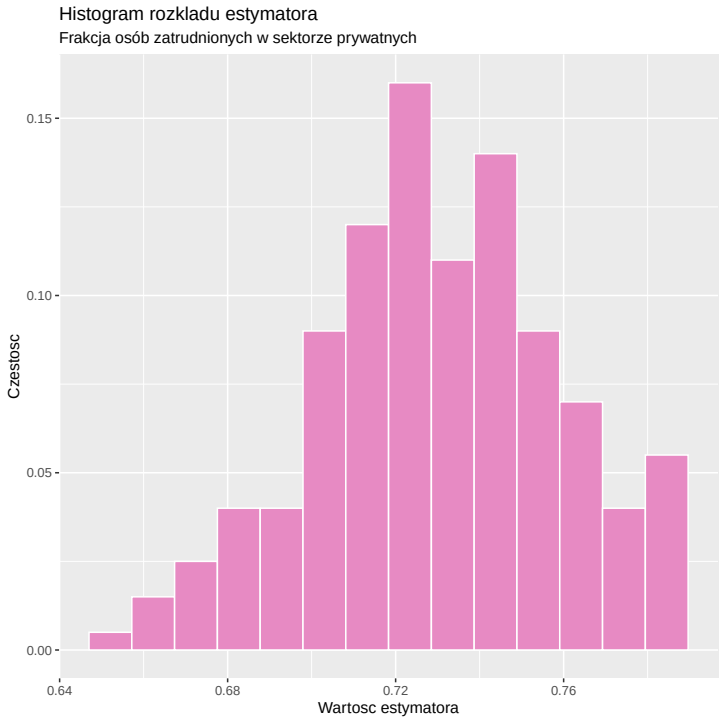
```
## Srednia szerokosc przedzialu:  0.1368086
## Czestosc zawierania:  0.965
```



Dla frakcji kobiet najczęściej występującą wartością estymatora jest około 0.45. Przedział jest odrobinę szerszy niż dla zmiennej p_W, co skutkuje wyższą częstością zawierania.

Dla frakcji p_P:

```
## Srednia szerokosc przedzialu:  0.121649
## Czestosc zawierania:  0.965
```



Podobnie jak dla poprzedniej frakcji, częstość zawierania jest odrobinę większa niż 0.95, ale częstość występowania wartości estymatora zbliżonej do 0.73 jest najwyższa.

1.3 Klasyczny przybliżony przedział ufności

Klasyczny (przybliżony) przedział ufności konstruujemy biorąc $\hat{p} = \frac{Y}{n}$ za środek przedziału. Budując SE jako $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Przedział ufności na poziomie α:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} SE_{\hat{p}} \tag{4}$$

Korzystamy z wcześniej wyliczonych sum dla Y. Funkcja do wyznaczenia klasycznego przedziału ufności:

```
classic_conf <- function(Y) {
  n <- 200
  p_hat <- Y/n
  z <- qnorm(0.975)
  SE <- sqrt(p_hat*(1 - p_hat)/n)
  left <- p_hat - z*SE
  right <- p_hat + z*SE

  cat("Przedzial ufnosci: [", left, ",", right, "]")
}
```

Dla p_W:

```
## Przedzial ufnosci: [ 0.2317969 , 0.3582031 ]
```

Dla p_K:

```
## Przedzial ufnosci: [ 0.3662928 , 0.5037072 ]
```

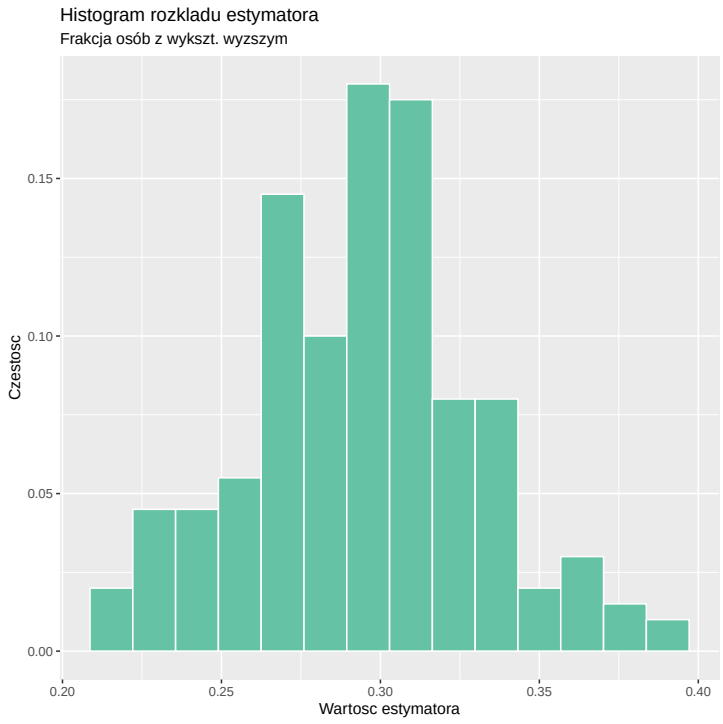
Dla p_P

```
## Przedzial ufnosci: [ 0.6631174 , 0.7868826 ]
```

Wyznaczone przedziały ufności są dość podobne do tych wyliczonych metodą Agrestiego-Coulla. Powtarzając eksperyment polegający na 200-krotnym losowaniu próby sprawdzimy rozkład estymatorów oraz szerokość przedziałów i częstość pokrycia rzeczywistej wartości frakcji przez przedział.

Histogram rozkładu estymatora oraz wyniki dla klasycznego przedziału ufności dla frakcji osób z wykształceniem wyższym:

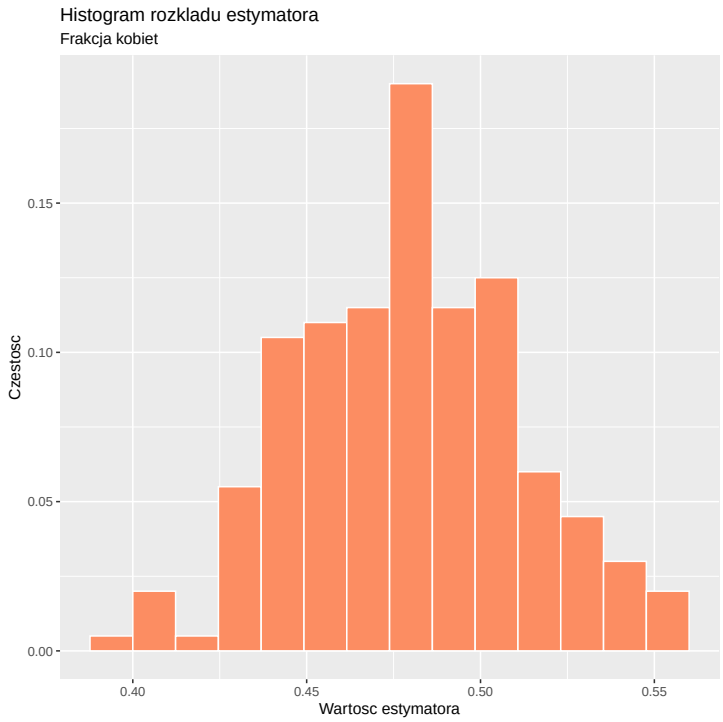
```
## Srednia szerokosc przedzialu:  0.1258481
## Czestosc zawierania:  0.9
```



Średnia szerokość przedziału jest dość podobna jak w przypadku przedziału Agrestiego-Coulla, jednakże częstość zawierania wynosi 0.9, a nie 0.95. To już różnica o cały poziom istotności. Najczęściej występujące wartości estymatora wynoszą około 0.3, co jest zgodne z rzeczywistą wartością.

Histogram rozkładu estymatora oraz wyniki dla klasycznego przedziału ufności dla frakcji kobiet:

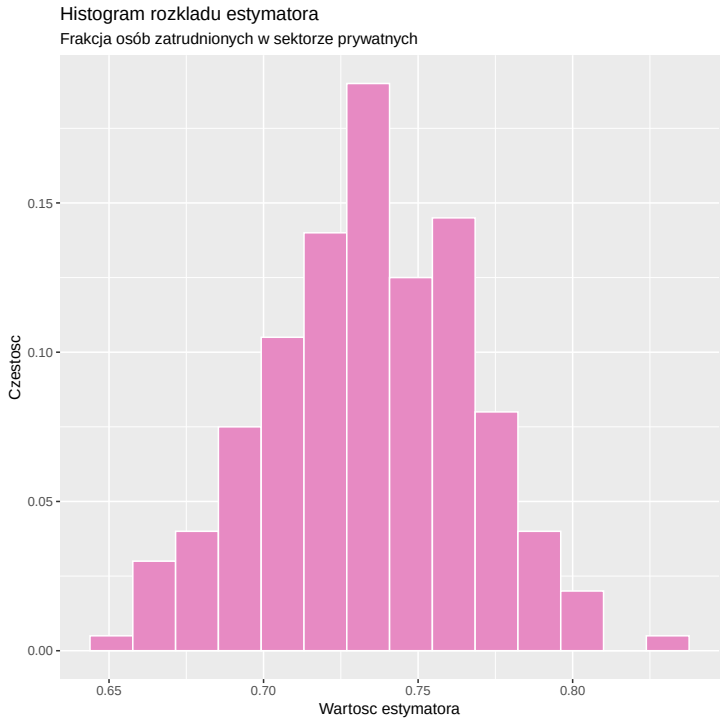
```
## Srednia szerokosc przedzialu:  0.1381903
## Czestosc zawierania:  0.965
```



Częstość zawierania jest taka sama jak w przypadku przedziału Agrestiego-Coulla, jednakże najczęściej występującą wartością estymatora jest wartość wynosząca około 0.47, a poprzednio było to około 0.45, jednakże są to wartości zbliżone do rzeczywistej wielkości frakcji kobiet. Częstość zawierania jest taka sama jak dla przedziału Agrestiego-Coulla.

Histogram rozkładu estymatora oraz wyniki dla klasycznego przedziału ufności dla frakcji osób pracujących w sektorze prywatnym:

```
## Srednia szerokosc przedzialu:  0.1219706
## Czestosc zawierania:  0.93
```



Częstość zawierania rzeczywistej wartości jest niższa niż dla przedziału Agrestiego–Coulla, ale najczęściej występująca wartość estymatora jest podobna.

2 Zadanie 2

Symulujemy 100-krotny rzut monetą i konstruujemy 95% przedział ufności dla frakcji orłów dwoma metodami: Agrestiego-Coulla i klasyczną. Następnie porównamy wynik z teoretycznym wyliczeniem dla symetrycznej monety opartym o przybliżenie rozkładu Bernoulliego rozkładem normalnym.

2.1 Przedział ufności Agrestiego-Coulla

Przedział budujemy zgodnie ze wzorem omówionym w sekcji 1.1.

```
## Przedzial ufnosci: [ 0.375102 , 0.5671176 ]
```

Możemy sprawdzić poprawność wyznaczonego przedziału przy pomocy funkcji `BinomCI` z biblioteki `DescTools`.

```
##            est   lwr.ci    upr.ci
## [1,] 0.4711098 0.375102 0.5671176
```

Wyznaczone przez funkcję granice przedziału są takie same jak te wyznaczone przez nas wprost ze wzoru.

2.2 Klasyczny przedział ufności

Klasyczny przedział ufności konstruujemy zgodnie ze wzorem w sekcji 1.3.

```
## Przedzial ufnosci: [ 0.3721784 , 0.5678216 ]
```

Sprawdzamy poprawność wyznaczonego przedziału:

```
##       est    lwr.ci    upr.ci
## [1,] 0.47 0.3721784 0.5678216
```

Przedział został poprawnie wyznaczony, co potwierdzają wyniki zwrócone przez funkcję `BinomCI`.

2.3 Teoretyczny przedział ufności

Korzystamy z faktu, że rozkład dwumianowy zmiennej Y (liczba sukcesów) przybliżamy dla dużych n rozkładem normalnym z parametrami $\mu = np$ i $\sigma = \sqrt{np(1-p)}$. Zatem zmienna $\hat{p} = \frac{Y}{n}$ ma wartość oczekiwaną $\mu = p$ oraz odchylenie standardowe $\sigma = \sqrt{\frac{p(1-p)}{n}}$, stąd $\hat{p}$ ma rozkład $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Stąd teoretyczny przedział ufności to:

$$\mu \pm Z_{\frac{\alpha}{2}}\sigma, \tag{5}$$

gdzie $\mu = p$.

```
## Przedzial ufnosci: [ 0.4020018 , 0.5979982 ]
```

2.4 Porównanie

Możemy zauważyć, że przedziały faktycznie różnią się od siebie. Porównamy ich szerokości.

1. Szerokość przedział ufności Agrestiego-Coulla: 0.192
2. Szerokość klasycznego przedziału ufności: 0.196
3. Szerokość teoretycznego przedziału ufności: 0.196

Najwęższy jest przedział Agrestiego-Coulla, jednakże wyniki są do siebie zbliżone.

3 Podsumowanie i wnioski

W raporcie sprawdziliśmy zachowanie trzech różnych przedziałów ufności dla frakcji sukcesów, rozumianych zarówno dosłownie (zadanie 2), jak i jako występowanie pewnej cechy w populacji (zadanie 1).

Potwierdziliśmy doświadczalnie tezę postawioną na wykładzie, mówiącą o tym, że klasyczny przedział ufności źle się zachowuje, gdy liczba sukcesów Y jest bliska 0 lub bliska n. Widać to szczególnie w zadaniu 1, gdzie jedynym wynikiem takim samym dla przedziału Agrestiego-Coulla i klasycznego było pokrycie rzeczywistej wartości przez przedział dla frakcji kobiet, które stanowią około połowy całej populacji zebranej w zbiorze danych. Zarówno dla małej frakcji osób z wykształceniem wyższym jak i dla dużej frakcji osób pracujących w sektorze prywatnym zastosowanie przedziału klasycznego skutkowało

mniejszym niż 95% prawdopodobieństwem pokrycia rzeczywistej wartości przez przedział ufności.

W zadaniu 2 zobaczyliśmy, że przedział ufności Agrestiego-Coulla jest węższy od przedziałów klasycznego i teoretycznego, co mogłoby skutkować bardziej dokładnymi wynikami badań przy zastosowaniu takiej konstrukcji przedziału ufności.

4 Adnotacja

We wszystkich wzorach zastosowałam zapis dotyczący kwantyli z rozkładu normalnego rzędu $\frac{\alpha}{2}$. W kodzie w R używam kwantyli rzędu $1 - \frac{\alpha}{2}$.