PSP – Raport 3

Nela Tomaszewicz

Maj 2020

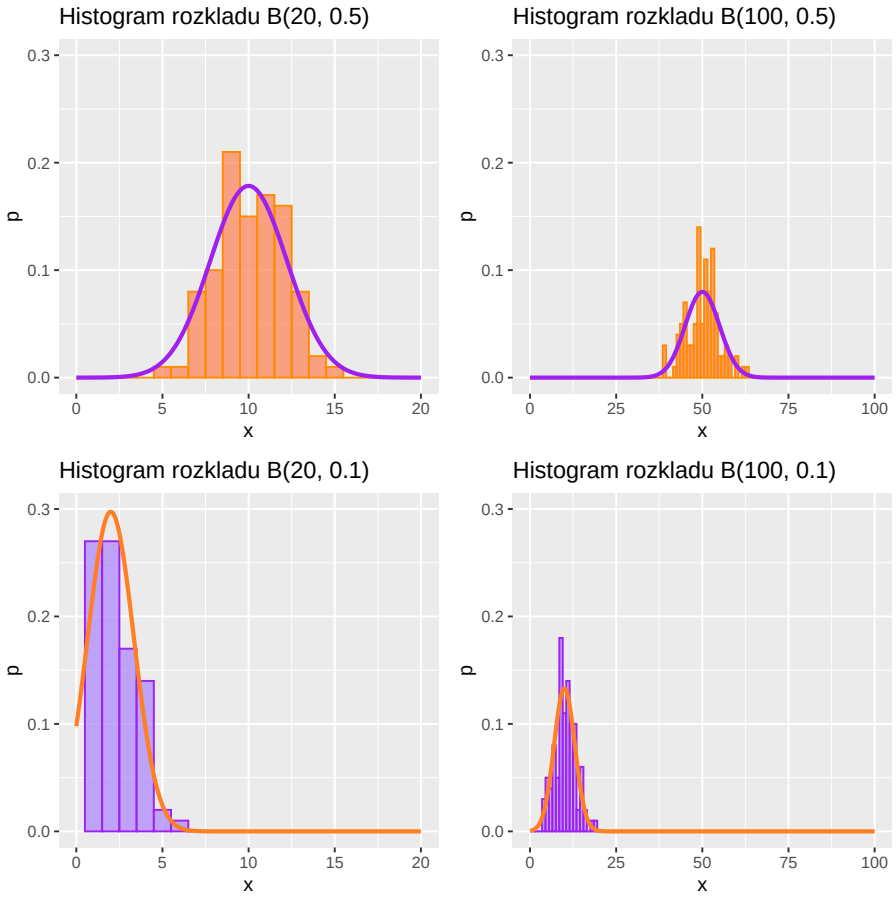
1 Zadanie 1

W zadaniu konstruujemy histogram rozkładu dwumianowego dla podanych zestawów parametrów, a następnie dorysowujemy wykres gęstości dla rozkładu normalnego z odpowiednio wyliczonymi wartościami μ i σ. Parametryzacja rozkładu dwumianowego $B(n, p)$ do rozkładu normalnego jest następująca:

$$\mu = np,$$

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

gdzie n oznacza liczbę prób, a p to prawdopodobieństwo sukcesu. Obserwacje generujemy funkcją `rbinom`, która przyjmuje 3 parametry: p i *size* (podane w zadaniu) oraz n oznaczające liczbę obserwacji do wygenerowania. Przyjmujemy, że dla każdego rozkładu losujemy $n = 100$ obserwacji. Wygenerowane histogramy wyglądają następująco.



Każdy z histogramów ma odpowiednio przeskalowane osie X i Y. Oś X ma wartości od 0 do n, natomiast oś Y przyjmuje wartości od 0 do 0.3. Szerokość binów w każdym z histogramów jest równa 1.

Najbliżej rozkładu normalnego są: pierwszy histogram dla $B(20, 0.5)$ oraz drugi histogram dla $B(100, 0.5)$. Funkcja gęstości jest tutaj najbardziej symetryczna. Jeśli chodzi o skośność, dla obydwu histogramów jest ona dość podobna. Dla pierwszego wynosi około -0.03, natomiast dla drugiego 0.04. Obydwie wartości można przybliżyć do 0. Dwa pozostałe histogramy dla $B(20, 0.1)$ i $B(100, 0.1)$ są mocno prawoskośne, z czego histogram dla 20 obserwacji jest mocniej skośny niż histogram dla 100 obserwacji.

1.1 Podsumowanie i wnioski

Zadanie pokazuje, że nawet gdy liczba prób jest duża, ale prawdopodobieństwo sukcesu nie jest równe 0.5, rozkład nie będzie nawet w przybliżeniu normalny (4. histogram). Najlepsze dopasowanie rozkładu dostajemy dla $p = 0.5$ i w tym przypadku liczba prób nie ma znaczenia. Najdalej od rozkładu normalnego jest histogram dla $n = 20$ i $p = 0.1$. Dodatkowo sprawdziłam, czy skośność dla $n > 100$ i $p = 0.5$ jest mniejsza. Owszem, skośność dla $n = 200$ jest równa około 0.2, zatem można wyciągnąć wniosek, że rozkład dwumianowy jest tym bliższy normalnemu im wielkość próby n jest większa, a prawdopodobieństwo sukcesu bliższe 0.5.

Warto zaznaczyć, że ogólnie histogram nie jest najlepszą formą do przedstawienia rozkładu dyskretnego, ponieważ wartości są kategoriami i nie powinno ich się zawierać w przedziałach przeznaczonych dla zmiennych ciągłych.

2 Zadanie 2

Rozważamy standardowy rozkład normalny $N(0, 1)$, dla którego wartości oczekiwanej skonstruujemy przedział ufności na poziomie 95%. Wzór na przedział ufności dla wartości oczekiwanej rozkładu normalnego (μ) przy znanym odchyleniu standardowym to:

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right], \tag{1}$$

gdzie n to liczba obserwacji, $\bar{X}$ oznacza średnią z obserwacji, α oznacza poziom istotności, $z_{1-\frac{\alpha}{2}}$ to kwantyl rzędu $1 - \frac{\alpha}{2}$ z rozkładu normalnego $N(0, 1)$, natomiast σ to odchylenie standardowe populacji.

2.1 Podpunkt (a)

Generujemy 100-elementową próbę z rozkładu $N(0, 1)$ i konstruujemy przedział ufności na poziomie 95% dla wartości oczekiwanej przy pomocy wzoru (1). Kod oraz wynik są następujące.

```
set.seed(1)
# kwantyl z rozkladu normalnego
z <- qnorm(0.975)
# licznosc proby
n <- 100
```

```r
# odchylenie std
sigma <- 1

#generowanie proby
prob <- rnorm(n, 0, 1)
#lewe ograniczenie przedzialu
left <- mean(prob) - z*(sigma/sqrt(n))
#prawe ograniczenie przedzialu
right <- mean(prob) + z*(sigma/sqrt(n))
#przedzial ufnosci
paste("[",round(left,2), ",",round(right,2), "]")

## [1] "[ -0.09 , 0.3 ]"
```

Znamy rzeczywistą wartość oczekiwaną, więc możemy stwierdzić, czy zawiera się w wyznaczonym przedziale. W tym przypadku tak.

2.2 Podpunkt (b)

Doświadczenie z podpunktu (a) powtarzamy 1000 razy i wyznaczamy jak często konstruowane przedziały ufności zawierają rzeczywistą wartość oczekiwaną. Z definicji przedziału ufności wynik powinien być zbliżony do wartości poziomu ufności, czyli 95%. Funkcja służąca do wykonania zadania zostanie przedstawiona poniżej.

```r
rep_conf_int <- function(n) {
  sum <- 0
  sigma <- 1
  z <- qnorm(0.975)
  for(i in 1:1000) {
    prob <- rnorm(n, 0, 1)
    left <- mean(prob) - z*(sigma/sqrt(n))
    right <- mean(prob) + z*(sigma/sqrt(n))
    if(0 >= left & 0 <= right) {
      sum = sum + 1
    }
  }
  return(paste("Czestosc: ",sum/1000))
}
```

Funkcja przyjmuje interesującą nas liczbę elementów, które chcemy wygenerować w każdym doświadczeniu. Zwraca częstość zawierania rzeczywistej wartości oczekiwanej. Wywołanie dla $n = 100$ zwraca następującą wartość.

```r
rep_conf_int(100)

## [1] "Czestosc:  0.962"
```

Częstość zawierania przez przedziay ufności rzeczywistej wartości oczekiwanej wynosi w przybliżeniu 95%, co jest zgodne z początkowym przypuszczeniem.

2.3 Podpunkt (c)

Powtarzamy doświadczenia z punktów (a) i (b) dla próby 200-elementowej. Przedział ufności dla takiej próby to:

```
## [1] "[ -0.02 , 0.25 ]"
```

Już teraz możemy zauważyć, że jest on trochę węższy niż przedział dla $n = 100$ obserwacji. Szerokość dla $n = 200$ wynosi 0.27, natomiast dla $n = 100$ to 0.39. Prawdopodobieństwo pokrycia rzeczywistej wartości oczekiwanej oraz średnia szerokość przedziałów ufności została wyznaczona poprzez zmodyfikowaną funkcję z podpunktu (b).

```
rep_conf_int <- function(n) {
  sum <- 0
  sigma <- 1
  z <- qnorm(0.975)
  v_width <- c()

  for(i in 1:1000) {
    prob <- rnorm(n, 0, 1)
    left <- mean(prob) - z*(sigma/sqrt(n))
    right <- mean(prob) + z*(sigma/sqrt(n))
    if(0 >= left & 0 <= right) {
      sum = sum + 1
    }
    #policzenie szerokosci przedzialu
    width <- right - left
    #zapisanie wyliczonych szerokosci
    v_width <- c(v_width, width)
  }
  return(paste("Czestosc: ",sum/1000, ",", "Srednia szerokosc: ",
               round(mean(v_width), 2)))
}
```

Wywołania dla $n = 100$ i $n = 200$ są następujące:

```
rep_conf_int(100)

## [1] "Czestosc:  0.95 , Srednia szerokosc:  0.39"

rep_conf_int(200)

## [1] "Czestosc:  0.941 , Srednia szerokosc:  0.28"
```

Częstość pokrycia (prawdopodobieństwo zawierania) rzeczywistej wartości oczekiwanej w obydwu przypadkach jest zbliżona do 0.95, czyli 95%. Średnia szerokość przedziału ufności dla większej próby jest mniejsza, co wynika wprost ze wzoru (1), gdzie wielkość próby n znajduje się w mianowniku.

2.4 Podsumowanie i wnioski

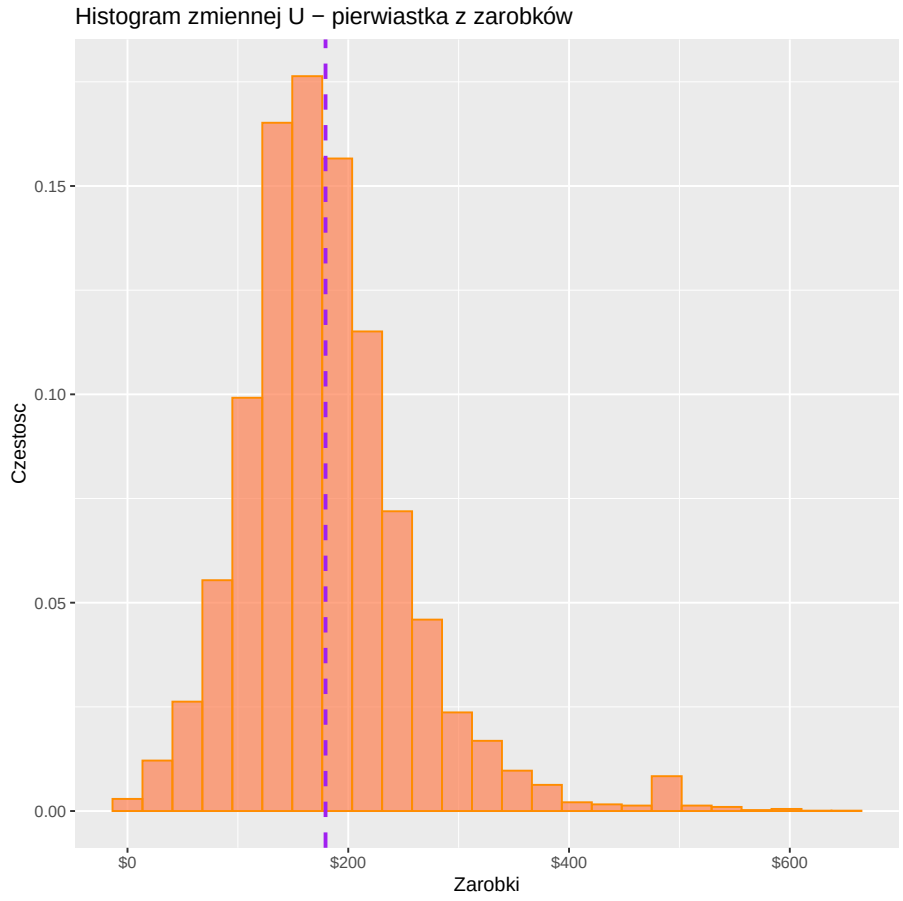
Im większa próba, tym przedział ufności dla wartości oczekiwanej jest węższy, co sprawia, że jesteśmy "bliżej" rzeczywistej wartości oczekiwanej dla całej populacji nawet gdy posługujemy się tym samym poziomem ufności. Wniosek jest dość logiczny: im większa próba, tym bardziej wiarygodne badania, ponieważ tym lepiej jesteśmy w stanie oszacować rzeczywistą wartość średnią.

3 Zadanie 3

W zadaniu ponownie korzystamy ze zbioru danych `income.txt` zawierającego dane dotyczące 55 899 pracowników z 3 sektorów zatrudnienia w USA. Dane zawierają także wiek pracowników, ich zarobki, płeć oraz wykształcenie.

3.1 Podpunkt (a)

Konstruujemy nową zmienną U będącą pierwiastkiem z dochodów (zarobków) pracowników, a następnie rysujemy jej histogram.



Rozkład zmiennej U nie jest normalny, co również było pokazane w Raporcie 2. Średnia wartość U to $\mu_U \approx 179.41$ (pokazana na wykresie przerywaną linią), natomiast średnia wartość dochodów pracowników to $\mu_D = 37864.61$

3.2 Podpunkt (b)

Losujemy 200-elementową próbę ze zbioru danych `income.txt`. Następnie wyznaczamy estymatory μ_U i μ_D będące średnią próbkową dla odpowiednio zmiennej U i D. Następnie na podstawie wylosowanej próby skonstruujemy 95% przedziały ufności i sprawdzimy, czy zawierają one rzeczywiste wartości średnie dla całej populacji wyznaczone w poprzednim podpunkcie.

Ponieważ nie znamy odchylenia standardowego, wzór na przedział ufności dla wartości średniej jest następujący:

$$\left[\bar{x} - t_{(1-\frac{\alpha}{2},n-1)}\frac{s}{\sqrt{n}}, \bar{x} + t_{(1-\frac{\alpha}{2},n-1)}\frac{s}{\sqrt{n}}\right], \tag{2}$$

gdzie $\bar{x}$ to średnia próbkowa, n to liczebność próby, $t_{(1-\frac{\alpha}{2},n-1)}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ z rozkładu t-Studenta z $n-1$ stopniami swobody, gdzie α to poziom istotności, a s jest próbkowym odchyleniem standardowym, czyli:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \tag{3}$$

Warto zaznaczyć, że powyższe wzory mogą być zaaplikowane do naszego przypadku na mocy Centralnego Twierdzenia Granicznego, mówiącego o tym, że dla dużego rozmiaru próby zmienna będąca średnią próbkową ma rozkład bliski normalnemu. Można zatem zadać pytanie: co to znaczy, że n (rozmiar próby) jest „duży"? Mówi się, że $n > 30$ jest wystarczające, aby zadziałało Centralne Twierdzenie Graniczne, jednakże jest to jedynie wiedza „empiryczna". Budujemy zatem *asymptotyczne* przedziały ufności, ponieważ przybliżamy rozkład normalny dla coraz większego rozmiaru próby.

Rozwiązanie zadania jest następujące.

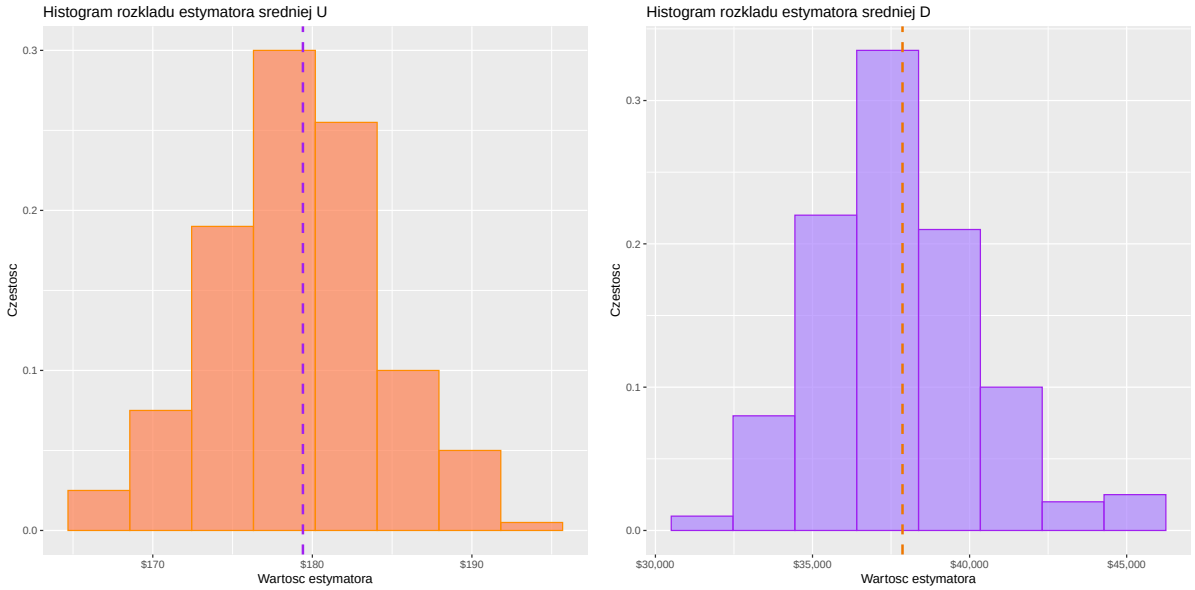
```
n <- 200
t <- qt(0.975, n-1)
# pobranie próbek
samp_inc <- sample(income$zarobki, n)
samp_inc_U <- sqrt(samp_inc)

#estymatory
est_U <- mean(samp_inc_U, na.rm=T)
est_D <- mean(samp_inc)
paste("Estymator U: ", round(est_U, 2))

## [1] "Estymator U:  172.72"

paste("Estymator D: ", round(est_D, 2))

## [1] "Estymator D:  34009.11"

#standardowy blad sredniej
SE_D <- sd(samp_inc, na.rm = T)/sqrt(n)
SE_U <- sd(samp_inc_U, na.rm = T)/sqrt(n)
```

```r
# przedzial dla U
left_U <- est_U - t * SE_U
right_U <- est_U + t * SE_U
paste("Przedzial ufnosci dla U: [",
      round(left_U, 2), ",", round(right_U,2), "]")

## [1] "Przedzial ufnosci dla U: [ 163.69 , 181.76 ]"

# przedzial dla D
left_D <- est_D - t*SE_D
right_D <- est_D + t*SE_D
paste("Przedzial ufnosci dla D: [",
      round(left_D, 2), ",", round(right_D,2), "]")

## [1] "Przedzial ufnosci dla D: [ 30202.35 , 37815.86 ]"
```

Przedziału ufności zawierają rzeczywiste parametry, czyli te wyliczone dla całej populacji, pomimo tego, że estymatory μ_U i μ_D różnią się od rzeczywistych wartości tego parametru.

3.3 Podpunkt (c)

Popdunkt (b) powtarzamy 200 razy i wyznaczamy histogramy rozkładu estymatorów dla μ_U i μ_D. Ponadto, naszym zadaniem jest wyznaczenie jak często przedziały ufności zawierały rzeczywistą wartość estymowanego parametru. Mając doświadczenie z poprzedniego zadania możemy postawić hipotezę, że częstość zawierania rzeczywistej wartości średniej przez przedział ufności powinna być zbliżona do 95%.



Przerywaną linią została zaznaczona rzeczywista wartość średnia dla zmiennych U i D. Wartość z przedziału, w którym średnia jest zawarta występował najczęściej. Dodatkowo, histogramy są zbliżone kształtem do rozkładu normalnego. Estymatorami parametrów μ_U i μ_D są po prostu średnie próbkowe, zatem można przypuścić, że histogramy przedstawiają działanie Centralnego Twierdzenia Granicznego.

Częstość zawierania rzeczywistej wartości μ_U i μ_D została wyznaczona przy pomocy analogicznej funkcji jak w zadaniu 2. Wyniki są następujące:

```
## [1] "Czestosc dla U:  0.935"
## [1] "Czestosc dla D:  0.945"
```

Częstości zawierania zbliżone są do 0.95. Budujemy przedziały asymptotyczne, stąd wartości te nie wynoszą równo 0.95.

3.4 Podsumowanie i wnioski

Zadanie pokazało, że pomimo tego, że rozkład badanych danych nie jest normalny, jesteśmy w stanie zbudować dobre przedziały ufności, czyli zawierające, z prawdopdobieństwem 0.95, średnią wartość badanej zmiennej dla całej populacji. Dodatkowo przetestowaliśmy działanie Centralnego Twierdzenia Granicznego, czyli jednego z najważniejszych twierdzeń w statystyce.

4 Zadanie 4

Korzystamy ze zbioru danych `grades.txt` zawierającego dane 78 uczniów pewnej szkoły w USA. Zakładamy, że ten zbiór jest prostą próbą losową z pewnej większej populacji. Skonstruujemy dwa przedziały ufności: dla średniego ilorazu inteligencji oraz dla średniego wyniku testu psychologicznego w tej populacji.

Nie znamy odchylenia standardowego dla populacji, więc aby wyznaczyć przedział ufności, skorzystamy ze wzoru (2). Znów wyznaczamy asymptotyczne przedziały ufności, ponieważ stosujemy przybliżenie przedziału ufności, które jest coraz lepsze im większa jest próba.

4.1 Asymptotyczny przedział ufności dla średniego poziomu inteligencji

Na podstawie próby losowej asymptotyczny przedział ufności na poziomie 95% dla średniego wyniku IQ wynosi:

```
## [1] "[ 105.95 , 111.89 ]"
```

4.2 Asymptotyczny przedział ufności dla średniego wyniku testu psychologicznego

Na podstawie próby losowej asymptotyczny przedział ufności na poziomie 95% dla średniego wyniku testu psychologicznego wynosi:

```
## [1] "[ 54.16 , 59.76 ]"
```

4.3 Podsumowanie i wnioski

Nie znamy wartości średniej badanych zmiennych dla całej populacji, dlatego nie możemy sprawdzić, czy asymptotyczne przedziały są poprawne. Mówimy tu oczywiście o zaweiraniu w przedziale ufności średniej jako wartości oczekiwanej danego parametru, a nie średniej próbkowej, która z definicji wzoru na przedział ufności (2) jest w nim zawarta.