Raport 2 - Doświadczalne sprawdzenie prawa wielkich liczb

Alicja Wiączkowska

2023-03-26

Contents

Cel

Celem raportu jest doświadcalne potwierdzenie poprawności prawa wielkich liczb poprzez porównanie średniej zarobków całej populacji oraz średnich dla poszczególnych prób różnych wielkości.

Prawo wielkich liczb

Według tego prawa gdy liczba niezależnych powtórzeń eksperymentu dąży do nieskoności, średnia z wyników tych doświadczeń dąży do wartości oczekiwanej pojedynczego eksperymentu.

Dla niezależnych zmiennych losowych $X_1,\ldots,X_n$ o jednakowym rozkładzie i wartości oczekiwanej μ prawo wielkoch liczb można zapisać:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \mu$$

Twierdzenie to oznacza, że im większa jest liczba niezależnych powtórzeń tego samego doświadczenia, tym lepiej średnia ich wyników przybliża wartość oczekiwaną.
Opis bazy danych
Analizowana baza **income.dat** zawiera wyniki ankiet dotyczących zarobków przeprowadzonych w roku 2000 przez *Bureau of Labor Statistics* na reprezentatywnej próbie mieszkańców USA. Zbiór ten pełni rolę całości badanej ppulacji.

Przygotowanie bazy danych do pracy z nimi

Odpowiedni plik został pobrany i zmodyfikowany. Kolumny: `Plec`, `Wyksztalcenie` i `Zarudnienie` zmieniono na typ factor oraz poprawiono w nich wartości zmieniając liczby na zakodowane nimi nazwy, aby ramka danych była bardziej czytelna.

Zawartość poszczególnych kolumn

- `Wiek` podany w latach.
- `Wyksztalcenie`: podstawowe, niepełne średnie, średnie, niepełne wyższe, wyższe (licencjat), wyższe (magisterium).
- `Plec`: mezczyzna, kobieta.
- `Zarobki` roczne podane w dolarach.
- `Zatrudnienie` - sektor zatrudnienia: sektor prywatny, sektor publiczny, samozatrudnienie.

Błędne dane

Z poniższego podsumowania kolumny `Zarobki` można odczytać, że wśród danych obecne są też wartości ujemne (np. minimalne zarobki wynoszą -24998.

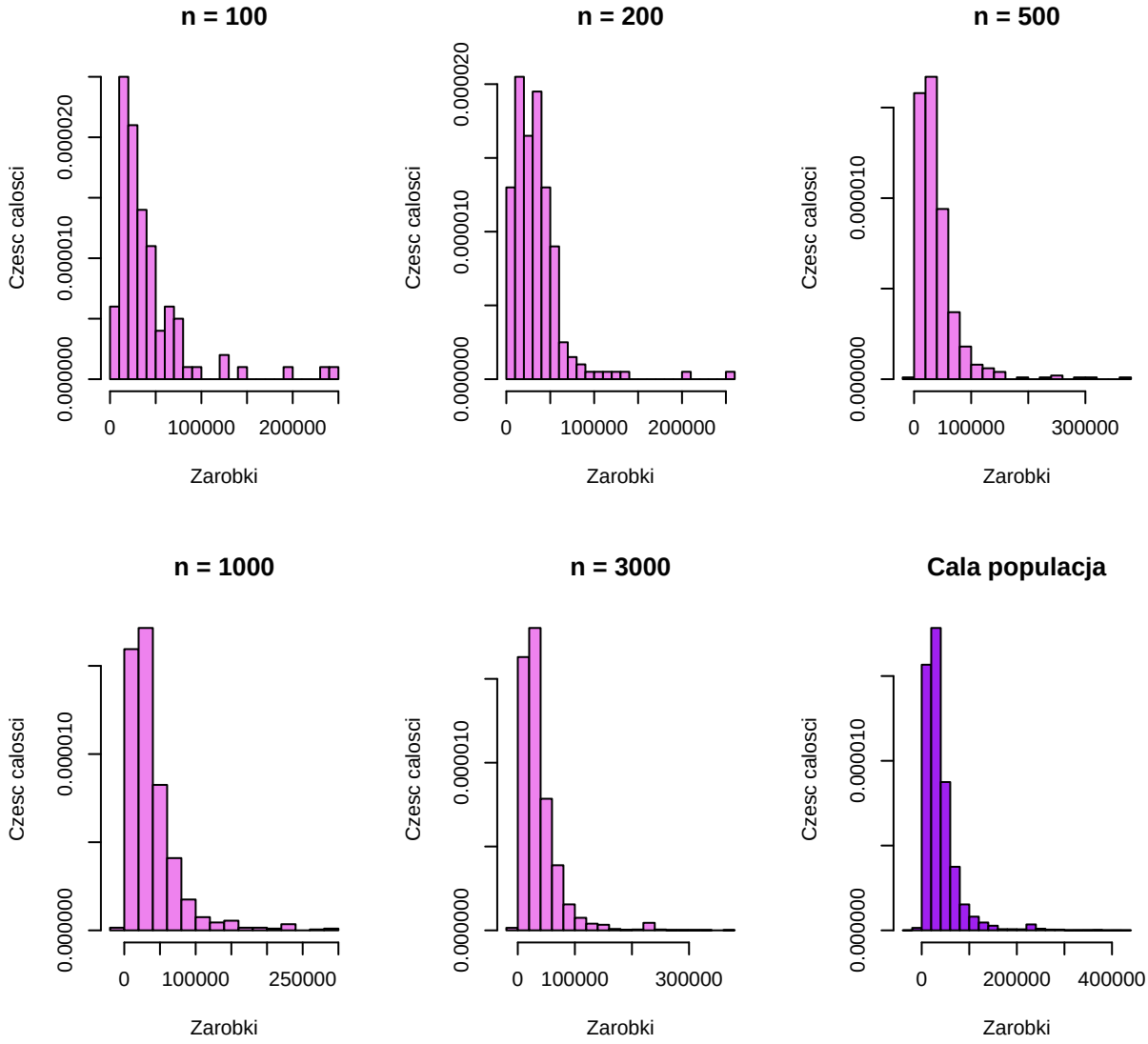
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -24998   17000   29717   37865   46504  425510
```

Uzyskanie dochodu poniżej zera jest możliwe tylko w przypadku posiadania własnej filrmy, co w USA traktowane jest jako samozatrudnienie (źródło: https://www.irs.gov/businesses/small-businesses-self-employed/self-employed-individuals-tax-center#SelfEmployed). Poniżej jednak z zestawienia zliczającego liczbe osób sektorach można odczytać, że również w sektorze prywatnym jest zatrudnione 0 osób. Potraktowao je jako błędnie wprowadzone dane i usunięto. Za zbędną uznano również kolumnę `L.p.`

```
##  sektor prywatny sektor publiczny samozatrudnienie
##                7                0              163
```

Histogramy i średnie zarobków

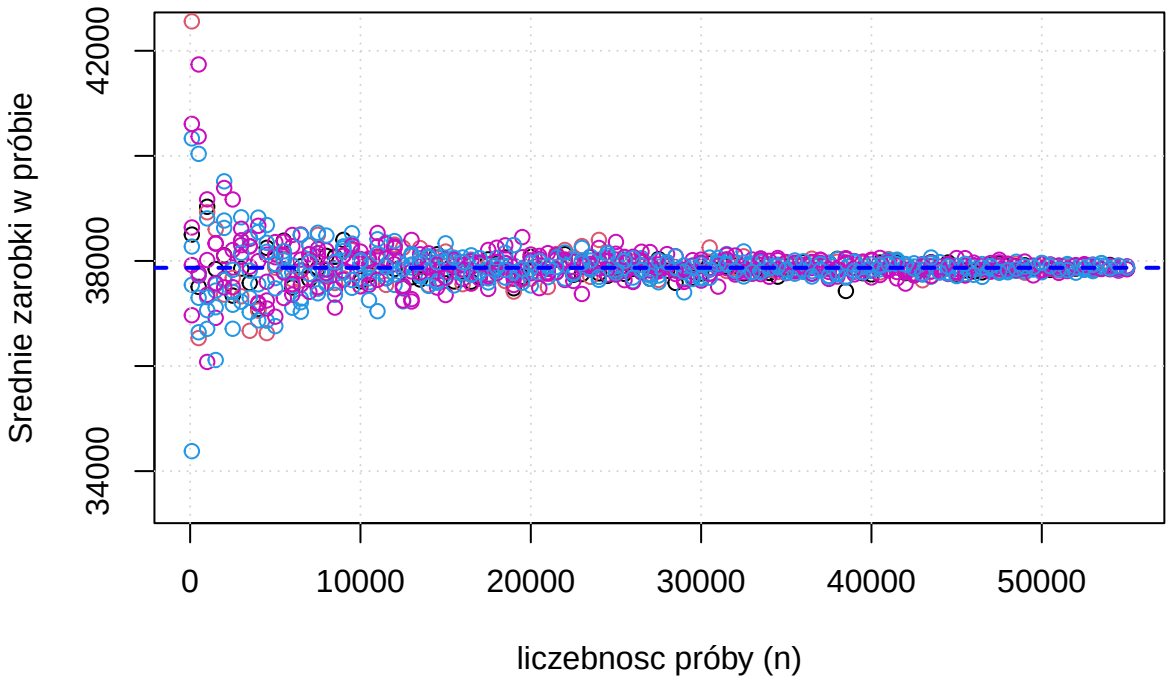
Poniżej przedstawiono histogramy zarobków dla losowo wybranych prób n-osobowych. Można zaobserwować, że wykresy wraz ze wzrostem n zbiegają kształtem do histogramu narysowanego dla całej populacji. Przy 20 przedziałach histgram już dla n=1000 (czyli 2% liczebności populacji) wygląda bardzo podobnie do histogramu całej populacji. Natomiast dla n=3000 (5%) wygląda niemal identycznie jak dla całości. Potwierdza to o fakt, że prawo wielkich liczb można stosować również w odniesieniu do rozkładów.



Poszczególne średnie zarobków dla prób z kórych sporządzone powyższych wykresów są zbliżone do średniej całej populacji wynoszącej 37869.73$.

	x
n = 100	41649.81
n = 200	33610.64
n = 500	38692.05
n = 1000	38850.93
n = 3000	37076.20

Po wielokrotnym sprawdzeniu średnich z losowej próby, można mieć pewność, że wraz ze wzrostem liczności próby, jej średnia zbliża się do wartości granicznej - średniej populacji (zaznaczonej przerywaną linią na wykresie). Zgodnie z oczekiwaniami, obserwacja potwierdza działanie prawa wielkich liczb.



Wielokrotne próbkowanie oraz teoretyczne odchylenie sandardowe

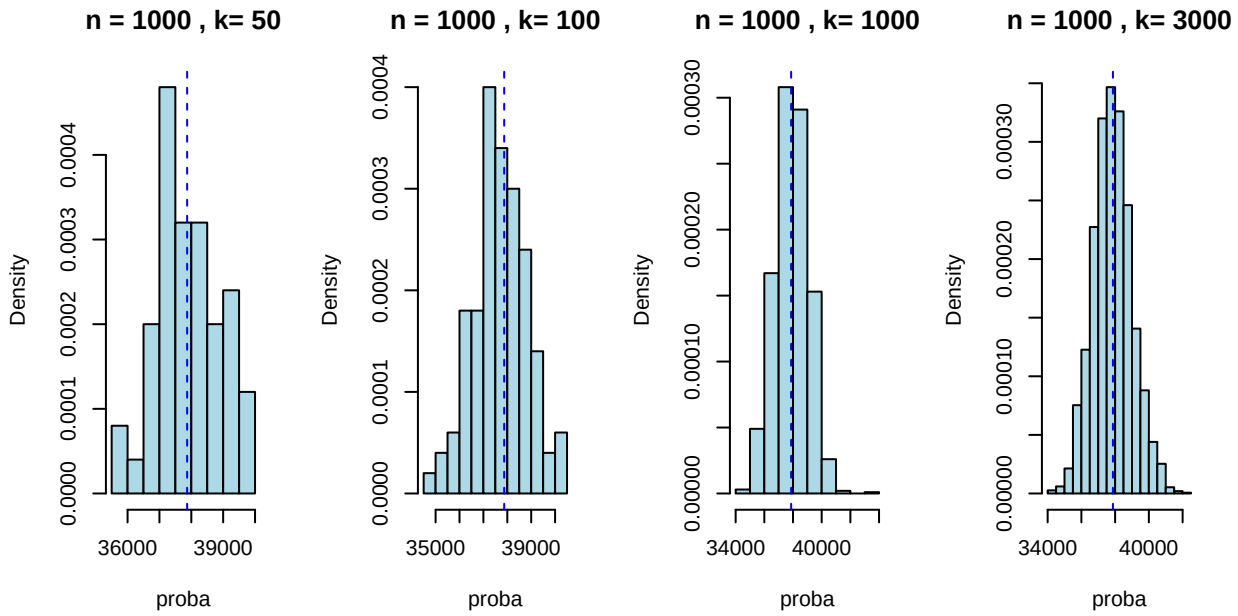
Teoretyczne średnie odchylenie standardowe dla obserwacji $X_i, ..., X_n$, z których każde X_i ma odchylenie standardowe σ można obliczyć w następujący sposób:

$$s = \sqrt{Var(\bar{X}_n)} = \sqrt{Var(\frac{1}{n}\sum_{i=1}^{n} X_i)} = \sqrt{\frac{1}{n^2} \cdot \sum_{i=1}^{n} Var(X_i)} = \sqrt{\frac{n \cdot \sigma^2}{n^2}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

W przypadku zarobków przyjmujemy σ równe odchyleniu standardowemu dla całej populacji.

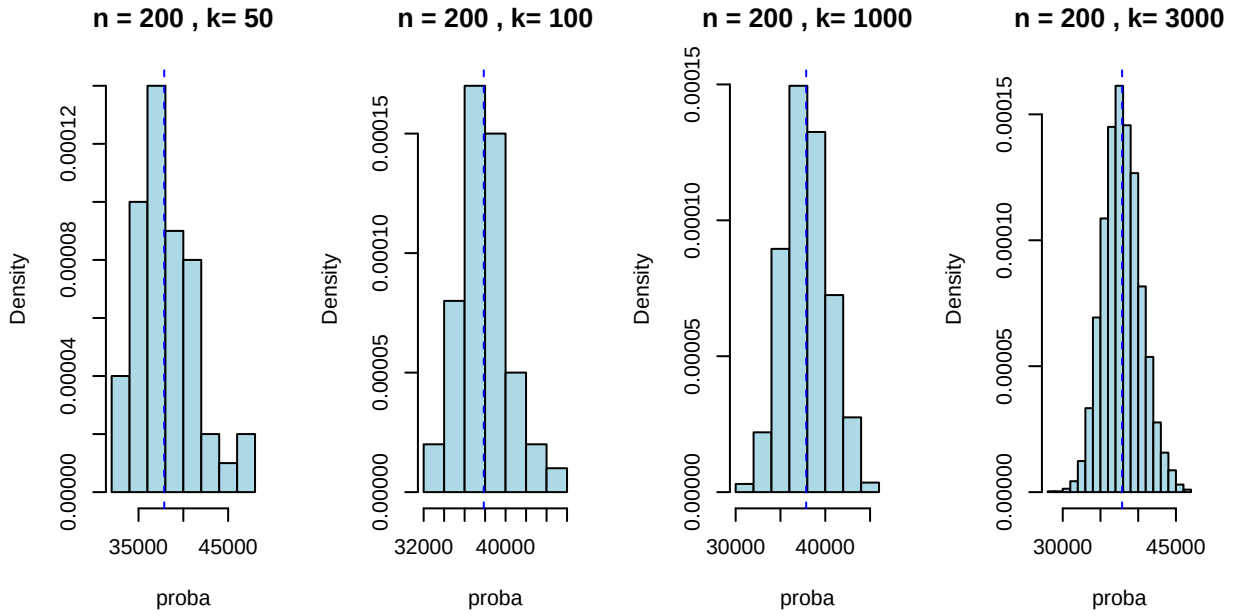
Przy ustalonej liczebności próby, histogramy średniej próbkowej wraz ze wzrostem k (liczby powtórzeń losowań) zbiegają do histogramu symetrycznego o środku zbliżonym do średnich zarobków populacji(wartość tą zaznaczono przerywaną linią). Ponadto z zestawień pod wykresami można odczytać, że wartości odchylenia standardowego wyliczone dla poszczególnych prób zbliżają się do wartości obliczonej teoretycznie wraz ze wzrostem k.

```
## [1] "dla n=1000"
```



k	doswiadczalnie	teoretycznie
50	986.9604	1143.397
100	1110.1235	1143.397
1000	1155.8194	1143.397
3000	1146.1278	1143.397

```
## [1] "dla n=200"
```



k	doswiadczalnie	teoretycznie
50	3160.855	2556.713
100	2433.957	2556.713
1000	2500.444	2556.713
3000	2507.969	2556.713