Zarobki mieszkańców USA w roku 2000

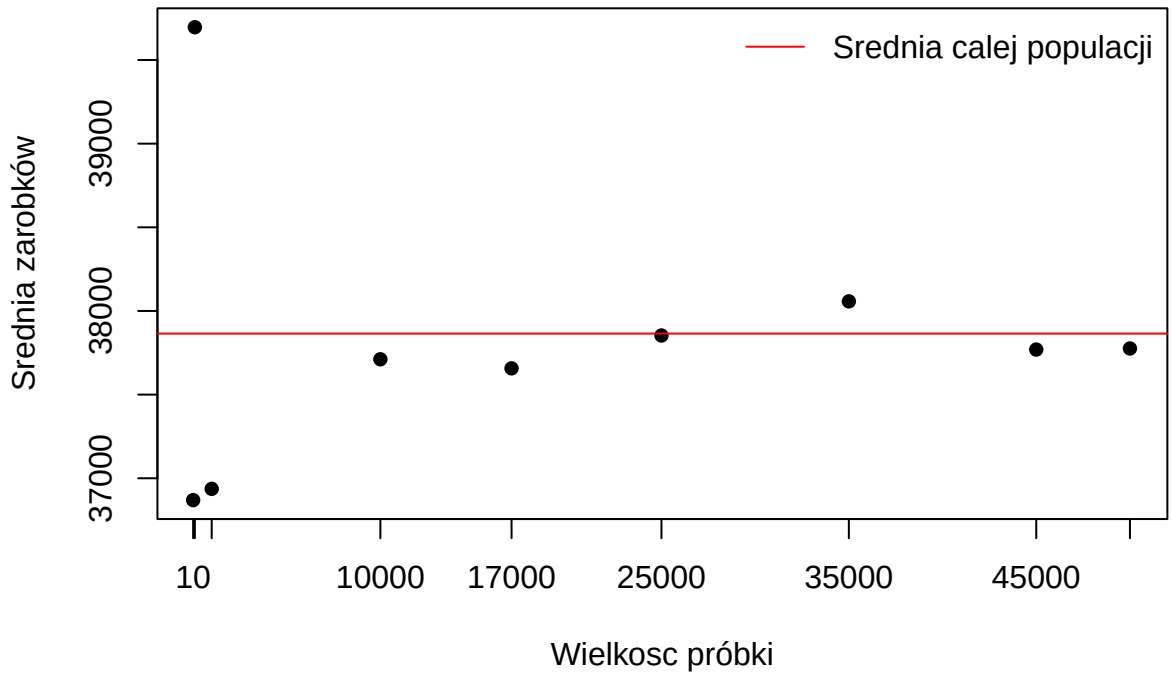
Magdalena Potok

2023-04-16

W tym raporcie wykorzystam dane zebrane w 2000 roku z ankiet przeprowadzonych przez Cureau of LAbor Statistics na reprezentatywnej próbie 55 899 mieszkańców. Zebrane dane zostały potraktowane jako informacje o całej populacji. Sprawdzę doświadczelnie Prawo wielkich liczb oraz szybkość zbieżności na trzy różne sposoby.

Porównanie średnich zarobków

Porównam teraz statystykę średnich zarobków całej populacji oraz średnich zarobków losowych prób różnych rozmiarów. Przedstawię to zróżnicowanie w postaci wykresu oraz ramki danych.



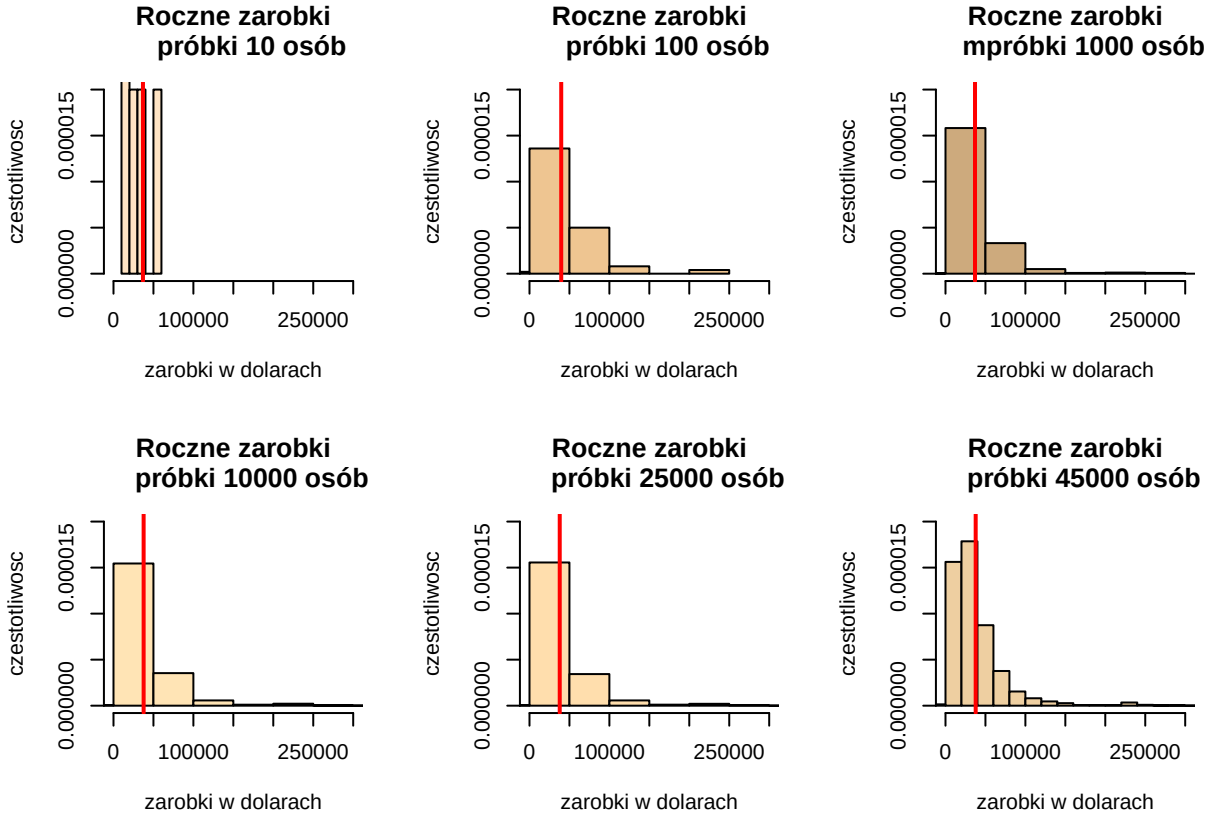
Średnia:

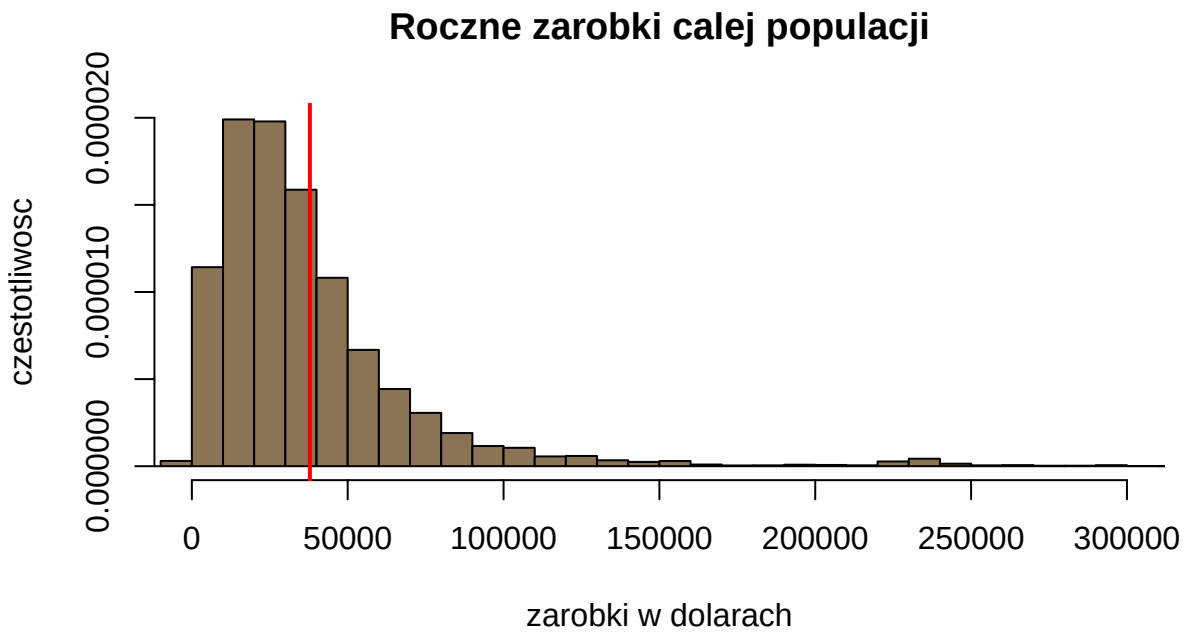
```
##   Całej populacji     Próbki 100 osób    Próbki 1000 osób   Próbki 10000 osób
##           37864.61            39695.77           36936.69           37711.15
## Próbki 25000 osób   Próbki 35000 osób   Próbki 45000 osób     Próbki 50000 osób
##           37853.33            38057.43           37769.35           37775.67
```

Prawo wielkich liczb to twierdzenie, które mówi nam, że w przypadku, gdy liczba powtórzeń eksperymentu jest wystarczająco duża, to częstość występowania zdarzenia będzie się mało różniła od prawdopodobieństwa tego zdarzenia.
W naszym przypadku prawdopodobieństwem zdarzenia będzie średnia całej populacji, a powtórzenia eksperymentu możemy rozumieć jako wybieranie coraz większego rozmiaru próbki badanych.
Z ramki danych oraz wykresu widać pewną prawidłowość, im wiekszą próbę wybierzemy, tym wynik jest faktycznie bliższy średniej wyliczonej z całej populacji. Losowanie wykonywane jest bez zwracania i jest to znaczący czynnik w naszych wynikach, ale widać, że już dla próbki 10000+ osób wynik jest bardzo blisko na wykresie granicy, którą jest średnia całej populacji, a więc już dla takiej wielkości próbki widać zbieżność.

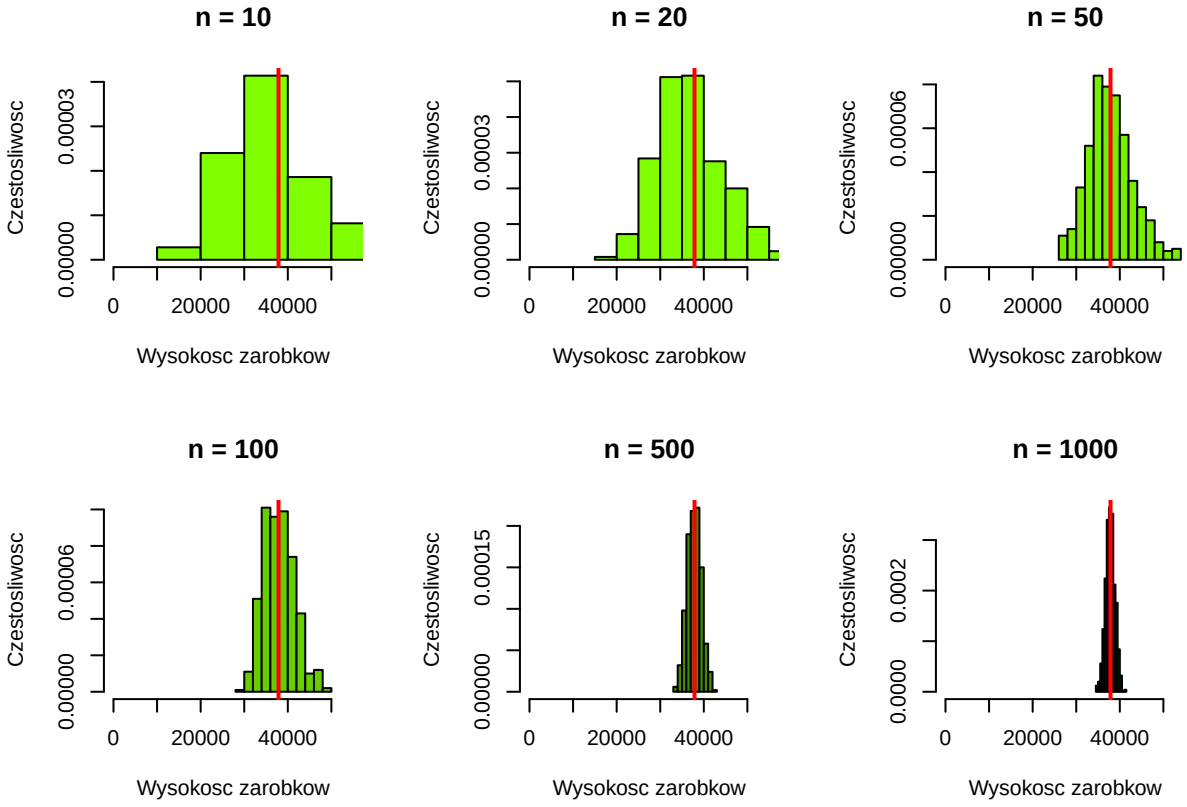
Porównananie histogramów zarobków

Poniżej znajdują się histogramy zarobków dla konkretnej wielkości próby. Kształt każdego z nich zależy od wylosowanych obserwacji, najbardziej odstający od reszty bywa wykres dla próbki 10 osób. Z racji tego, że jest to skrajnie mała ilość, jego kształt bywa przeróżny zależny od wylosowanych obserwacji). Jednak można zauważyć, że przy coraz większej ilości osób, histogramy przybierają kształt asymetryczny skośny w prawo. Ta informacja mówi nam, że znaczna większość mieszkańców USA w 2000 roku zarabiała poniżej średniej. Czerwonym kolorem zaznaczyłam średnią dla każdej wielkości próby, podkreśla ona przed chwilą wspomniany fakt.
Zgodnie z prawem o wielkich liczbach można zauważyć, że najbardziej przypominający histogram całej populacji jest histogram dla największej próby, czyli 45 000 osób. Widać jednak, że nawet wykres dla 10000 osób przypomina kształtem wykres całej populacji.





Histogram rozkladu sredniej probkowej zarobkow przy wielokrotnym probkowaniu



Powyżej przedstawiłam histogramy rozkłady średniej próbkowej zarobków, eksperyment został w każdym przypadku powtórzony 500 razy.
Można zauważyć, że wszystkie histogramy, przy rożnych wielkościach próbek, przypominają rozkład normalny. Już nawet dla próbki n = 100 widać duże podobieństwo z rozkładem normalnym, w tym histogramie słupki symetrycznie odległe od środka minimalnie między sobą się różnią, a to jest właśnie charakterystyczna cecha rozkładu normalnego. We wszystkich przypadkach środek histogramu leży w miejscu średniej calej populacji Zaznaczonej czerwonym kolorem na każdym histogramie.
Widać, że z rosnącą wielkością próbki histogramy zbiegają do wartości środkowej. Oznacza to, że z wielkością próbki histogram rozkładu, przy wielokrotnym próbkowaniu, zbiega do wartości oczekiwanej, co jest zgodne z Prawem wielkich liczb.

Poniżej przedstawię odchylenie standardowe tych rozkładów dla poszczególnych n.

```
##            10        20        50       100       500      1000
## sd   11434.17   8085.18   5113.52   3615.8   1617.04   1143.42
```

Jak widać im większe n, tym odchylenie standardowe maleje, co można było zauważyć już na histogramie, gdy słupki zbiegały i zaczeły skupiać się wokół średniej. Jest to kolejny przykład, w którym widać działanie Prawa wielkich liczb.

Sprawdzę teraz zasadę 68-95-99.7 dla każdego z powyższych rozkładów.

$$[\mu - \sigma, \mu + \sigma]$$

```
##              10   20     50    100  500   1000  10000
## ni+/-sigma   73   70   68.4   69.4   71   69.8   72.4
```

$$[\mu - 2\sigma, \mu + 2\sigma]$$

```
##               10     20     50     100    500   1000 10000
## ni+/-2sigma   95   96.6   95.2   95.4   95.8   97.2    96
```

$$[\mu - 3\sigma, \mu + 3\sigma]$$

```
##                 10     20     50     100    500   1000 10000
## ni+/-3sigma   99.2   99.4   99.6   99.6   99.8   99.8    99
```

Na podstawie powyższych wyliczeń można zauważyć, że rozkłady mniej więcej zachowują zasadę 68-95-99.7 rozkładu normalnego. Czyli można stwierdzić, że przybliżanie tego rozkładu rozkładem normalnym jest uzasadnione, co jest zgodne z Centralnym Twierdzeniem Granicznym.

Wnioski

Doświadczalnie sprawdziłam trzy razy Prawo wielkich liczb. Za każdym razem wraz ze wzrostem wielkości próbki, byłam coraz bliżej porządanego efektu - jakim była statystyka całej populacji. W każdym przypadku ta zbieżność była naprawdę szybka, już przy próbce wielkości 1000 można było zauważać podobieństwo zachowywania się jak próbka całej populacji.