

Zarobki mieszkańców USA w roku 2000

Magdalena Potok

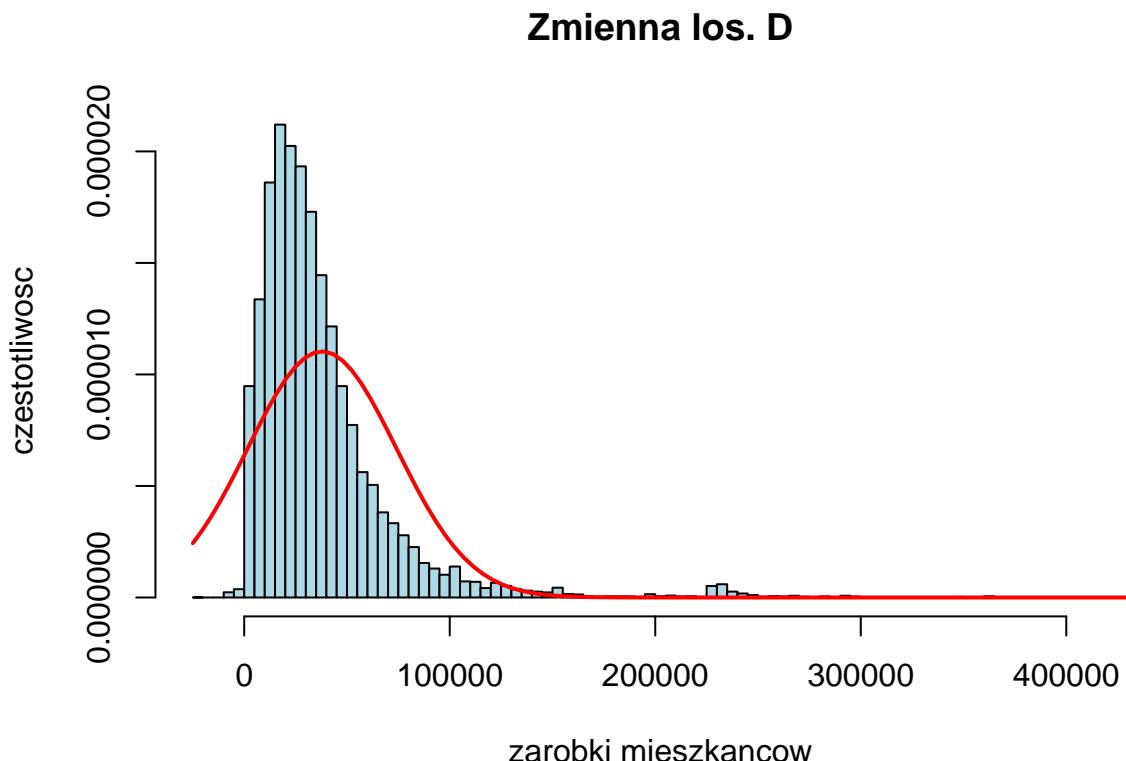
2023-04-16

W raporcie zostały wykorzystane dane dotyczące zarobków mieszkańców USA w 2000 roku. Dochód całej populacji został oznaczony literą D, a pierwiastek kwadratowy D zmienną U. Zbadam normalność obu zmiennych za pomocą pięciu różnych narzędzi.

Odczytanie normalności z histogramu

Aby zbadać normalność zmiennej z wykresu należy nałożyć na niego krzywą gęstości, która przedstawia rozkład prawdopodobieństwa zmiennej. Dla dostatecznie dużej wielkości próby histogram zmiennej o rozkładzie normalnym powinien przybliżać krzywą gęstości tego rozkładu.

Poniższy histogram przedstawia rozkład zmiennej D - czyli zarobków mieszkańców USA w 2000 roku, wielkość próby to 55 899. Na wykres została nałożona krzywa gęstości, dzięki której można zbadać normalność naszej zmiennej.



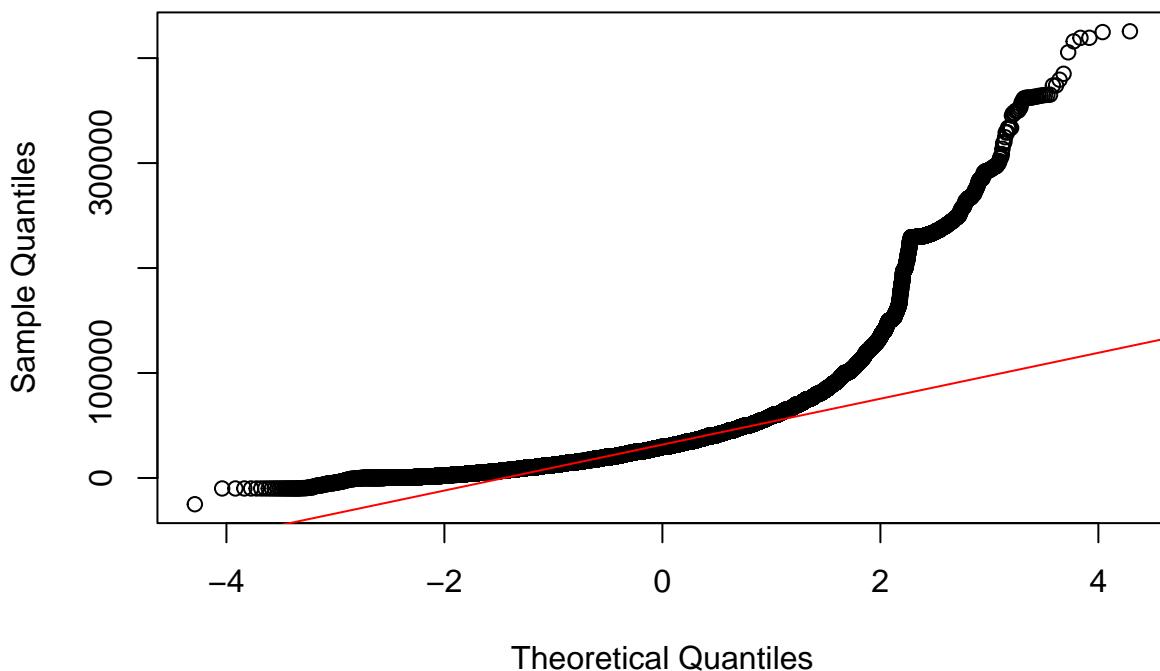
Już na pierwszy rzut oka widać, że histogram nie pokrywa się z krzywą gęstości i znacząca większość wykresu wystaje poza nią. Ta informacja oraz fakt, że histogram jest asymetryczny, skośny w prawo, wskazują nam,

że rozkład zmiennej D nie jest normalny.

Wykres Q-Q

Kolejnym sposobem zbadania normalności rozkładu jest wykres Q-Q. Taki wykres przedstawia punkty, w których osie x i y reprezentują odpowiednie kwantyle dla teoretycznego rozkładu i dla badanego rozkładu. Jeśli punkty na wykresie układają się wzduż prostej, to można przypuszczać, że rozkład jest zgodny z rozkładem teoretycznym. Prosta która przedstawia rozkład teoretyczny, jest wyznaczana na podstawie średniej i odchylenia standardowego zmiennej.

Wykres QQ zmiennej D



Powyżej widać wykres Q-Q dla badanego rozkładu - D. Kolorem czerwonym została dodana prosta przedstawiająca rozkład teoretyczny. Widać, że punkty nie układają się wzduż tej prostej, co oznacza, że rozkład zmiennej różni się od teoretycznego rozkładu prawdopodobieństwa. Z wykresu można odczytać, że to zróżnicowanie jest naprawdę duże. Co sugeruje nam, że podana zmienna nie ma rozkładu normalnego.

Reguła 68% - 95% - 99.7%

Następnym sposobem zbadania normalności rozkładu jest reguła 68% - 95% - 99.7%. Polega ona na tym, że sprawdzamy jaki procent obserwacji znajduje się w przedziałach:

- jednego odchylenia standardowego poniżej średniej do jednego odchylenia standardowego powyżej średniej,
- dwóch odchyleń standardowych poniżej średniej do dwóch odchyleń standardowych powyżej średniej,
- trzech odchyleń standardowych poniżej średniej do trzech odchyleń standardowych powyżej średniej.

Zbadanie normalności rozkładu polega na porównaniu uzyskanych procentów z odpowiednio 68% - 95% - 99.7% - są to oczekiwane wartości dla rozkładu normalnego. Jeśli te procenty są zbliżone do wartości teoretycznych, to możemy wnioskować, że rozkład jest normalny.

Poniżej przedstawię uzyskane procenty dla badanej zmiennej D.

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

```
## [1] 88.73325
```

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

```
## [1] 96.42391
```

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

```
## [1] 98.04111
```

Wyniki sugerują nam, że rozkład zmiennej D może nie być normalny, ponieważ obserwowane wartości procentowe znacznie różnią się od wartości teoretycznych dla rozkładu normalnego. Dla jednego sigma wynik jest aż o 20% wyższy, co wskazuje na to, że więcej danych znajduje się z tego przedziału, niż jest to oczekiwane dla rozkładu normalnego. Wynik 2 sigma jest zbliżony do oczekiwanej wartości, jednak dla 3 sigma wyszło mniej niż 99.7%

Shapiro-Wilk test

Następnym narzędziem do zbadania normalności rozkładu D będzie Sharpio-Wilk test. Mierzy on stopień odchylenia rozkładu badanej zmiennej od teoretycznego rozkładu normalnego. Im wartość statystyki testowej większa, tym bardziej dane są zgodne z rozkładem normalnym.

Niestety test Sharpio-Wilka ma ograniczenia i działa tylko dla prób mniejszych niż 5000, wobec tego wylosowałam 5000 obserwacji ze zmiennej D i przeprowadziłam dla nich test.

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  D1  
## W = 0.69451, p-value < 0.0000000000000022
```

Interpretacja wyniku testu zależy od przyjętego poziomu istotności alfa. Standardowo przyjmuje się 0,05. Zauważmy, że p-wartość jest bardzo mała, mniejsza niż ustalony poziom istotności, świadczy to o tym, że dane nie pochodzą z rozkładu normalnego.

Kolmogorov-Smirnov test

Następnie zbadam normalność zmiennej za pomocą testu K-S. Tym razem nie ma ograniczenia wielkości próby badanej zmiennej, wobec tego test zostanie przeprowadzony dla wszystkich obserwacji.

```
##  
##  Asymptotic one-sample Kolmogorov-Smirnov test  
##  
## data:  unique(D)  
## D = 0.99501, p-value < 0.0000000000000022  
## alternative hypothesis: two-sided
```

Ponownie jak w przypadku testu S-W patrzymy na wynik p-wartości. Wartość jest znacznie mniejsza od 0,05. Świadczy to o tym, że zmienna D nie ma rozkładu normalnego.

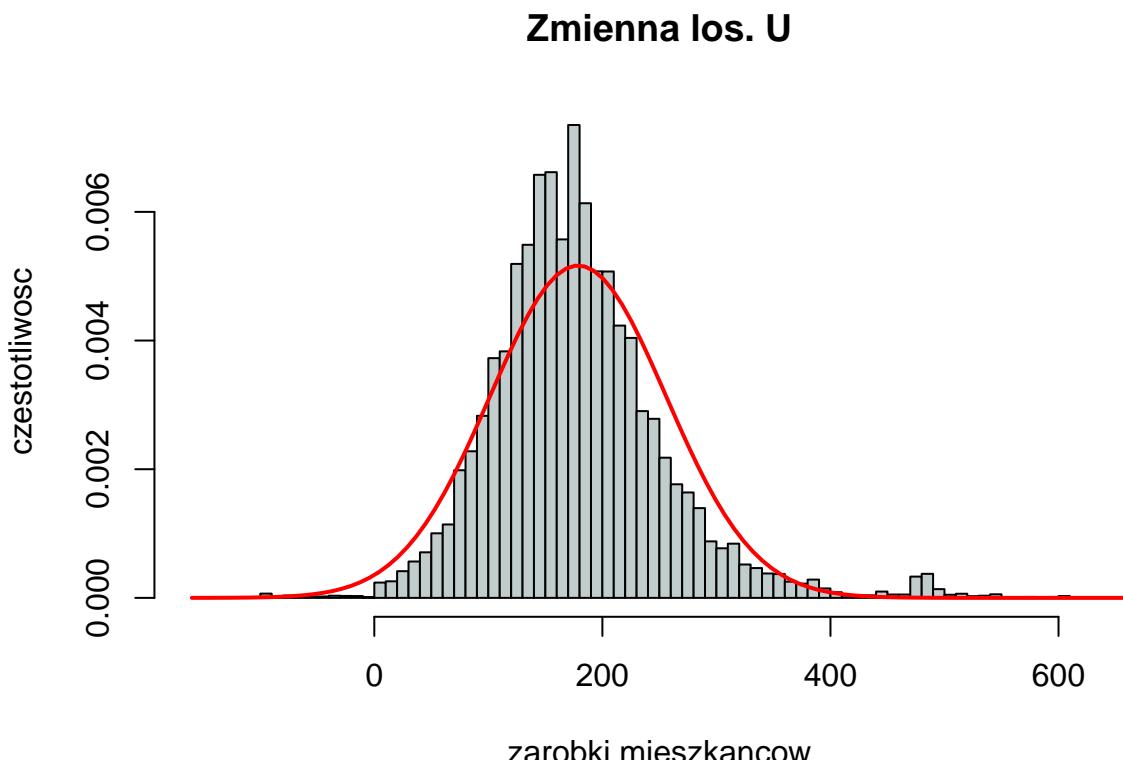
Badanie normalności zmiennej U

Jako zmienną U w przypadku gdy $D \geq 0$ przyjmuję pierwiastek kwadratowy z D. Natomiast gdy $D < 0$ będzie to pierwiastek kwadratowy z wartości bezwzględnej z D ze zmienionym znakiem.

$$\text{Prościej zapisując : } U = \sqrt{|D|} \cdot \text{sign}(D)$$

Zbadam teraz normalność zmiennej U używając tych smajch narzędzi, co dla zmiennej D.

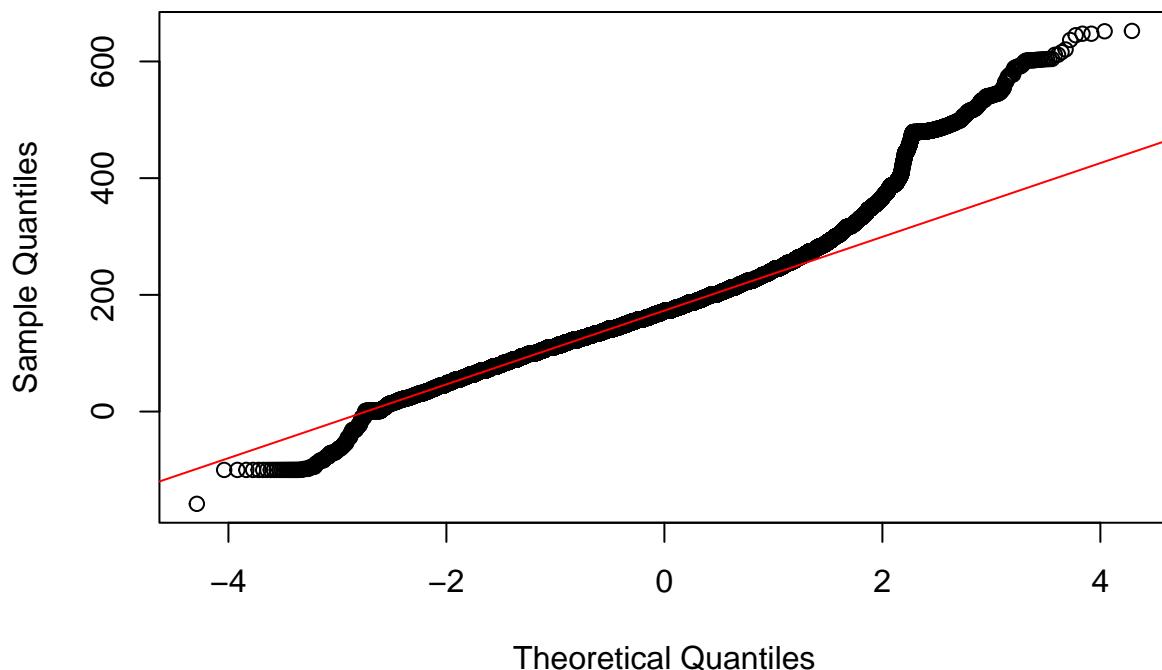
Odczytanie normalności z histogramu



Zarówno histogram jak i krzywa gęstości mają symetryczny i lekko skośny w prawo kształt. Zdecydowanie większa część histogramu znajduje się pod krzywą gęstości, co sugeruje nam, że badany rozkład jest bliski rozkładowi normalnemu.

Wykres Q-Q

Wykres QQ zmiennej U



Widac znaczącą różnicę między wykresem K-K dla zmiennej D, a dla zmiennej U. Tutaj znacznie większa ilość punktów układa się w prostą linię. Tak naprawdę jedynie krańce odstają od prostej linii, co może wskazywać na tzw. ‘ciężkie ogony’. Dane mają skrajne wartości, jednak w większości zachowują się tak, jak rozkład normalny. W takim przypadku ciężko potwierdzić, jak i zaprzeczyć tezie o normalności zmiennej U, warto zastosować inne narzędzia.

Reguła 68% - 95% - 99.7%

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

```
## [1] 75.58811
```

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

```
## [1] 95.53838
```

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

```
## [1] 98.34523
```

Zastosowanie tej metody, podobnie tak jak dla zmiennej D, nie daje nam oczekiwanych (jak dla rozkładu normalnego) wyników. Pierwszy różni się o 7%, drugi jest prawidłowy, natomiast trzeci różni się prawie o 1,5%. Z dwóch ostatnich metod można wysnuć wnioski, mówiące nam, że dane mają skrajne wartości, które zaburzają normalność badanego rozkładu.

Shapiro-Wilk test

```
##  
## Shapiro-Wilk normality test  
##  
## data: U1  
## W = 0.93698, p-value < 0.00000000000000022
```

Wynik testu Sharpia-Wilk'a dla wylosowanych 5000 obserwacji ze zmiennej U jest podobny do wyniku zmiennej D. Informuje nas o tym, że rozkład zmiennej U nie jest normalny.

Kolmogorov-Smirnov test

```
##  
## Asymptotic one-sample Kolmogorov-Smirnov test  
##  
## data: unique(U)  
## D = 0.99479, p-value < 0.00000000000000022  
## alternative hypothesis: two-sided
```

W przypadku K-S testu jest dokładnie tak samo, p-wartość zdecydowanie jest mniejsza od 0,05, co wskazuje na brak normalności.

Prawdopodobieństwo oraz wartości oczekiwane obu zmiennych

Wartość oczekiwana zmiennej D: $\mu_D = 37864.61$

Wartość oczekiwana zmiennej U: $\mu_U = 178.69$

Wartość oczekiwana zmiennej U do kwadratu: $(\mu_U)^2 = 31928.46$

Podniesiona wartość oczekiwana zmiennej U do kwadratu jest bliska wartości oczekianej D. Zmienna U jest bezpośrednio wyliczana jako pierwiastek ze zmiennej D, dlatego te wartości oczekiwane są podobne. Wobec czego możemy aproksymować wartość oczekowaną oraz odchylenie standardowe zmiennej D poprzez zmienną U. Użyję tego przybliżenia, aby policzyć prawdopodobieństwo szukane w zadaniu, przekształcając wzór w następujący sposób:

$$\begin{aligned} P(|D - \mu_D| \leq k \cdot \sigma_D) &= \\ P(|D - (\mu_U)^2| \leq k \cdot (\sigma_U)^2) &= \\ P(-k \cdot (\sigma_U)^2 + (\mu_U)^2 \leq D \leq k \cdot (\sigma_U)^2 + (\mu_U)^2) &= \\ P(D \leq k \cdot (\sigma_U)^2 + (\mu_U)^2) - P(D \leq -k \cdot (\sigma_U)^2 + (\mu_U)^2) & \end{aligned}$$

```
##      prawd.dla.zmiennej.D  prawd.przyblizone.  
## k = 1      0.682689492137086  0.682689492137086  
## k = 2      0.954499736103642  0.954499736103642  
## k = 3      0.99730020393674   0.99730020393674  
## k = 4      0.999936657516334  0.999936657516334
```

Z ramki danych można wyczytać, że prawdopodobieństwo wyliczone dla zmiennej D oraz przy pomocy przybliżenia wartości średniej oraz odchylenia standardowego tej zmiennej używając $(\mu_U)^2$ i $(\sigma_U)^2$ jest identyczne. Co oznacza, że znając rozkład zmiennej U, jesteśmy w stanie wyliczyć $P(|D - \mu_D| \leq k \cdot \sigma_D)$.