Politechnika Wrocławska
Wydział Matematyki

Skład grupy:	Agata Sobczak 268873
	Katarzyna Kudełko 268762
Prowadzący laboratorium:	dr inż. Rafał Połoczański
Prowadzący wykład:	dr hab. inż. Krzysztof Burnecki

Statystyka stosowana
Raport 1.

Analiza wybranych danych rzeczywistych z wykorzystaniem metod statystyki opisowej

Wiek zawarcia pierwszego małżeństwa kobiet z USA

WROCŁAW 7.5.2023

Spis treści

1 Wstęp

1.1 Informacje o danych

Do raportu zostały użyte dane ankietowe o wieku zawarcia pierwszego małżeństwa kobiet mieszkających w USA, które odpowiedziały na National Survey of Family Growth (NSFG) przeprowadzone przez Centers for Desease Control and Prevention (CDC) w cyklu lat 2006-2010. Zbiór danych X to wiek kobiet. W próbie znajduje się 5534 wierszy danych.
Dane użyte w tym raporcie należą do NSFG, cdc.gov/nchs/nsfg/nsfg_2006_2010_puf.htm. Można pobrać je ze strony vincentarelbundock.github.io/Rdatasets/datasets.html (Age at first marriage of 5,534 US women).

1.2 Cel raportu

Celem raportu jest przeanalizowanie oraz zwizualizowanie danych dotyczących wieku zawarcia pierwszego małżeństwa kobiet USA na przestrzeni lat 2006-2010, a także identyfikacja pewnej tendencji w ówczesnym czasie. Wszystkie obliczenia zostały wykonane w języku Python przy użyciu bibliotek NumPy, scipy.stats, statistics oraz matplotlib.pyplot.

2 Podstawowe statystyki

Wszystkie wartości statystyczne zostały wyznaczone za pomocą programów stworzonych przez autorów, które wykorzystują estymatory nieobciążone lub wbudowane funkcje do ich obliczenia.

2.1 Tabele

	Dane
Mediana	23
Kwartyl rzędu $\frac{1}{4}$	20
Kwartyl rzędu $\frac{3}{4}$	26
Rozstęp miedzykwartylowy	6
Rozstęp	33
Wariancja	22.29
Odchylenie standardowe	4.72
Kurtoza	0.73
Współczynnik skośności	0.77
Współczynnik zmienności	0.20 %

Tabela 1: Tabela charakterystyki danych

Średnia	Dane
arytmetyczna	23.44
geometryczna	25.56
harmoniczna	22.99

Tabela 2: Tabela wyliczonych średnich dla danych

2.2 Informacje o statystykach

- **Mediana** - to wartość, która dzieli uporządkowany zbiór danych na dwie równe części, takie że połowa wartości jest mniejsza lub równa medianie, a druga połowa jest większa lub równa medianie.

- **Kwartyle** - to miary pozycyjne, podobnie jak mediana, które pozwalają określić pozycję lub wartość w zbiorze danych. Kwantyle dzielą uporządkowany zbiór danych na równe części, ale w przeciwieństwie do mediany, kwantyle dzielą zbiór na więcej niż dwie części.

$$F(x_p) = p$$

- **Rozstęp międzykwartylowy** - to miara zmienności danych, wyznaczana jako różnica między trzecim kwartylem a pierwszym kwartylem zbioru danych.

$$R_Q = Q_3 - Q_1$$

- **Rozstęp** - to prosta miara zmienności danych, wyznaczana jako różnica między największą i najmniejszą wartością w zbiorze danych.

$$R = X_{max} - X_{min}$$

- **Wariancja** - to miara zmienności danych, która określa, jak bardzo wartości w zbiorze danych rozróżniają się od średniej arytmetycznej.

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- **Odchylenie standardowe** - to pierwiastek kwadratowy z wariancji, określa, jak bardzo wartości w zbiorze danych rozróżniają się od średniej arytmetycznej.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- **Współczynnik zmienności** - to miara względnej zmienności, czyli stosunek odchylenia standardowego do średniej arytmetycznej w zbiorze danych.

$$\frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}}{\frac{1}{n}\sum\limits_{i=1}^{n} x_i} \cdot 100\%$$

- **Średnia arytmetyczna** - to średnia wartość w zbiorze danych, która jest obliczana jako suma wszystkich wartości w zbiorze podzielona przez liczbę tych wartości.

$$\overline{x_a} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Średnia geometryczna** - to średnia wartość, która jest obliczana jako pierwiastek n-tego stopnia z iloczynu wszystkich wartości w zbiorze.

$$\overline{x_g} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

- **Średnia harmoniczna** - to odwrotność średniej arytmetycznej odwrotności wartości w zbiorze.

$$\overline{x_h} = \frac{n}{\sum\limits_{i=1}^{n}\frac{1}{x_i}}$$

- **Kurtoza** - jest to względna miara koncentracji i spłaszczenia rozkładu. Określa rozmieszczenie i koncentrację wartości w pobliżu średniej.

$$K = \frac{m_4}{s^4}$$

gdzie:
m_4 - momement centralny rzędu czwartego,
s^2 - odchylenie standardowe podniesione do czwartej potęgi.

- **Współczynnik skośności** - o miara asymetrii rozkładu prawdopodobieństwa. Wartość skośności wskazuje na to, w którą stronę rozkład jest asymetryczny i jak bardzo.

$$A_Q = \frac{Q_1 + Q_3 - 2m}{Q_3 - Q_1}$$

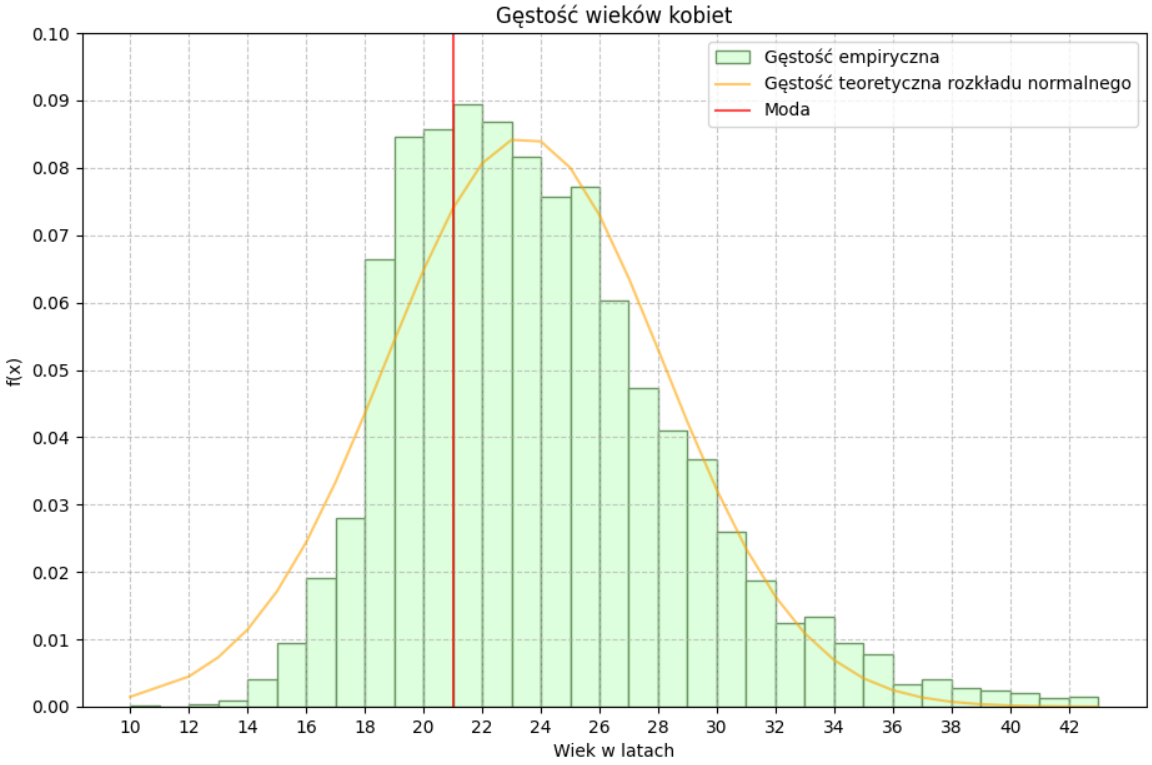
gdzie:
m - mediana

Dla zbioru danych współczynnik asymetrii i kurtoza jest dodatnia, co mówi o tym, że badany rozkład jest prawostronnie skośny i posiada bardziej strome szczyty i dłuższe ogony niż rozkład normalny.

3 Wizualizacja danych

Poniżej zostały przedstawione popularne sposoby wizualizacji danych, które pozwalają na zobrazowanie rozkładu wartości w danym zbiorze, takie jak: gęstość, dystrybuanta, wykres pudełkowy oraz kwantylowy, a także wykresy średniej ucinanej i winsorowskiej dla parametru k. Niektóre z nich porównano z ich odpowiednikami dla rozkładu normalnego, aby przekonać się, czy analizowane dane reprezentują rozkład normalny.

3.1 Wykres gęstości

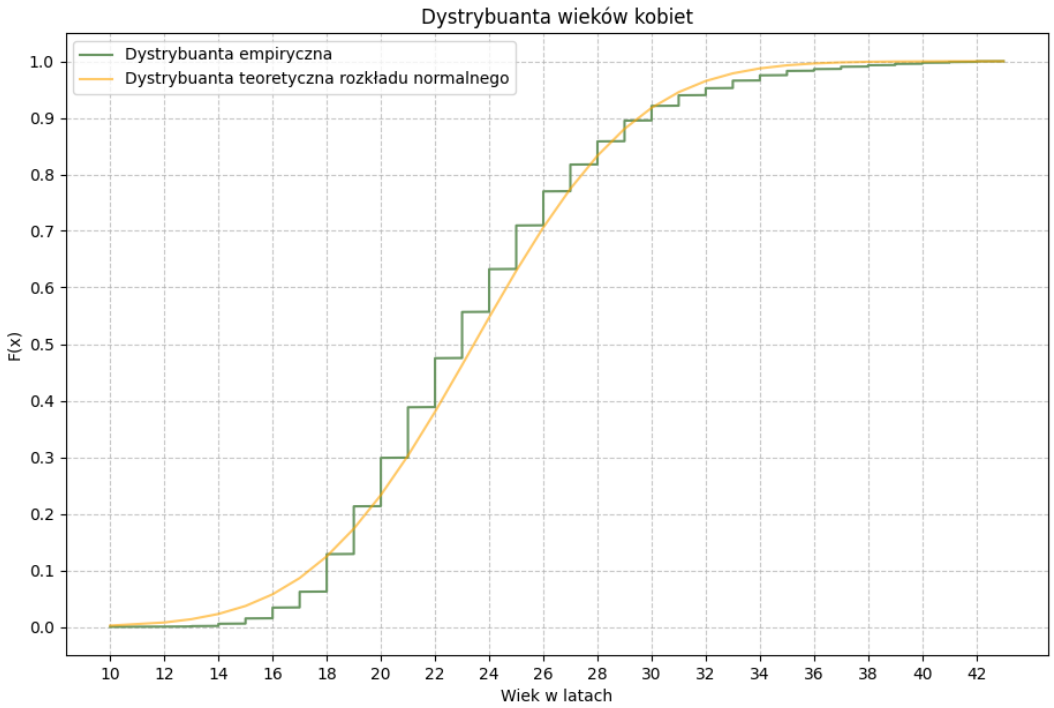


Rysunek 1: Gęstość wieków kobiet

Powyższy histogram reprezentuje częstość danych obserwacji w próbie. Jak widać, najwięcej obserwacji przypada na przedział $19 - 26$. Można również odczytać wartość najczęstszą - modę - równą 21. Na

rysunek naniesiono także wykres gęstości teoretycznej rozkładu normalnego, gdzie za parametry przyjęto odpowiednio średnią i odchylenie standardowe badanego zbioru danych. Obie gęstości zauważalnie się od siebie różnią, zatem można stwierdzić, że zbiór X nie jest próbą pochodzącą z rozkładu normalnego. Potwierdza to również moda - dla rozkładu normalnego wynosiłaby ona ok. $23 - 24$.

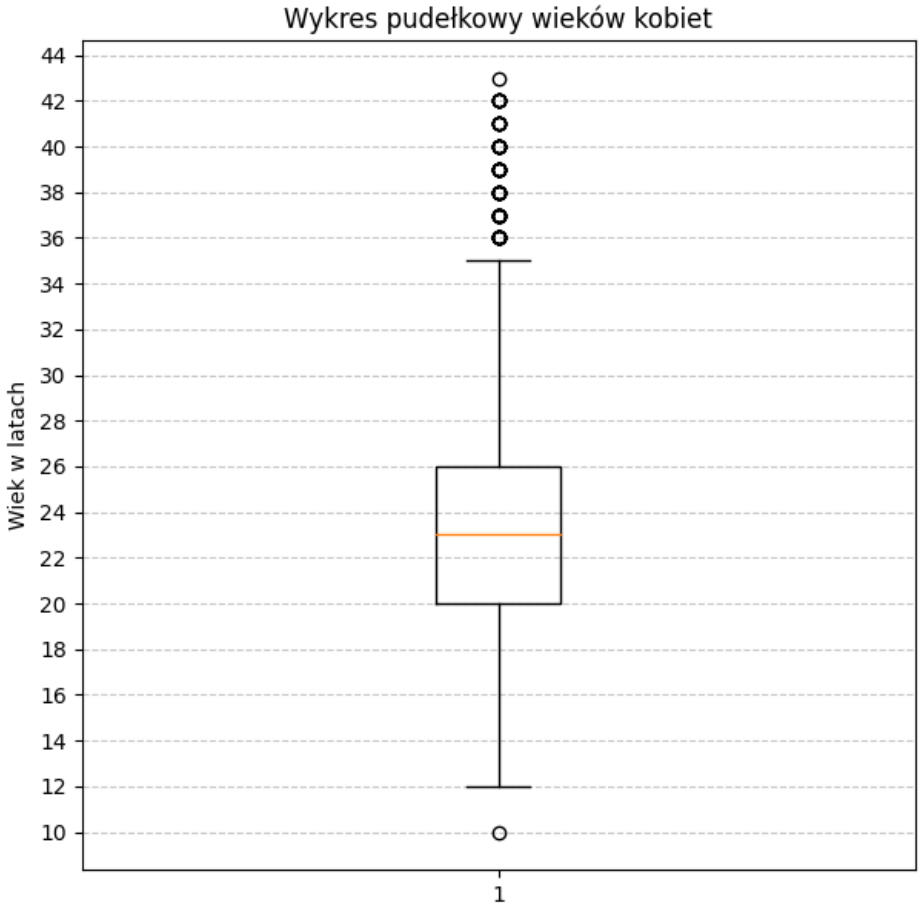
3.2 Wykres dystrybuanty



Rysunek 2: Dystrybuanta wieków kobiet

Wykres dystrybuanty umożliwia dokładne określenie, jak często wartości w danym zbiorze danych są mniejsze lub równe danej wartości. Dla przykładu, szansa na wartość ≤ 21 wynosi $\sim 0,4$, gdzie dla rozkładu normalnego byłoby to $\sim 0,3$. Widać również, że dystrybuanta zbioru X jest schodkowa - oznacza to, że rozkład jest dyskretny, co zgadza się ze zbiorem danych, zatem można stwierdzić, że dystrybuantę zaimplementowano poprawnie. Ponownie, oznacza to również, że analizowane dane nie pochodzą z rozkładu normalnego, gdyż jego dystrybuanta jest ciągła.

3.3 Wykres pudełkowy (Box plot)



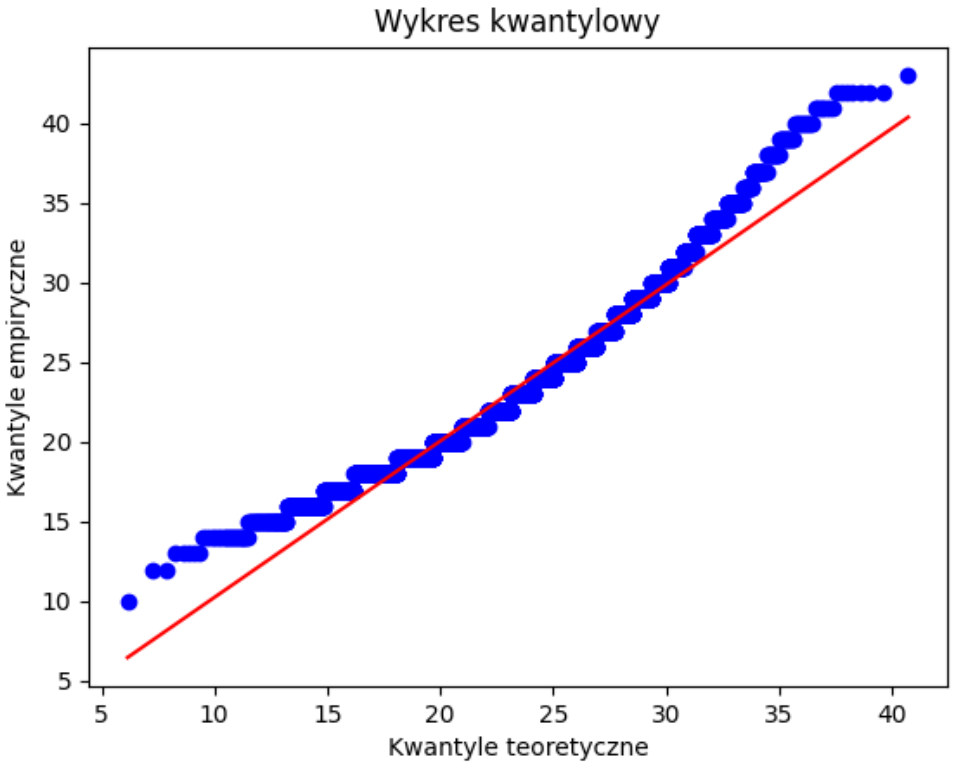
Rysunek 3: Wykres pudełkowy (Box plot) wieków kobiet

Z powyższego wykresu możliwe jest odczytanie niektórych wcześniej obliczanych wartości podstawowych statystyk. Wartości wykreślone na wykresie pudełkowym reprezentują kwartyle zbioru danych. Długość pudełka oznacza rozstęp ćwiartkowy kwartyli Q_1 *i* Q_3. Długość wąsów natomiast wyznacza się w następujący sposób:

- dł. dolnego wąsa $= Q_1 \breve{} 1.5 \cdot$ dł. pudełka
- dł. górnego wąsa $= Q_3 + 1.5 \cdot$ dł. pudełka

Pomarańczowa linia pozioma w pudełku oznacza medianę. Wartości odstające są oznaczone jako osobne punkty na wykresie. Odczytywane wartości są zgodne z wcześniej obliczonymi.

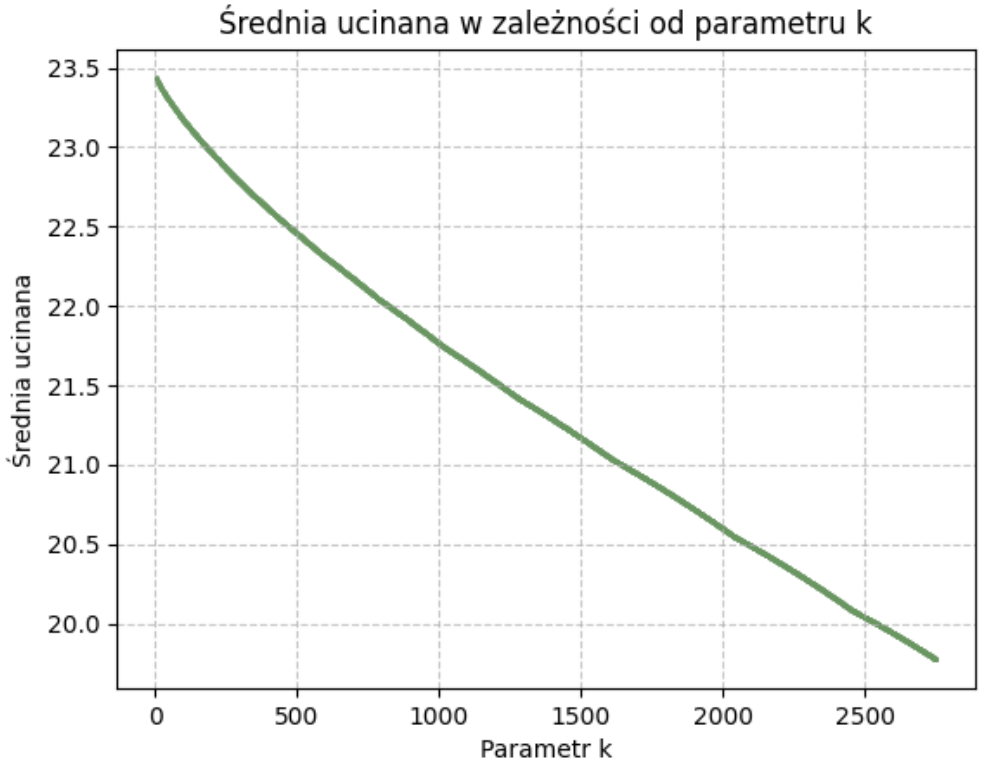
3.4 Wykres kwantylowy (Q-Q plot)



Rysunek 4: Wykres kwantylowy (Q-Q plot) wieków kobiet

Sprawdzono również jak wygląda wykres kwantylowy w porównaniu do rozkładu normalnego. Jak widać, wartości znacznie odbiegają od czerwonej linii oznaczającej rozkład normalny, przebiegającej przez środek wygenerowanego rysunku. Potwierdza to, że analizowane dane z pewnością nie pochodzą z rozkładu normalnego.

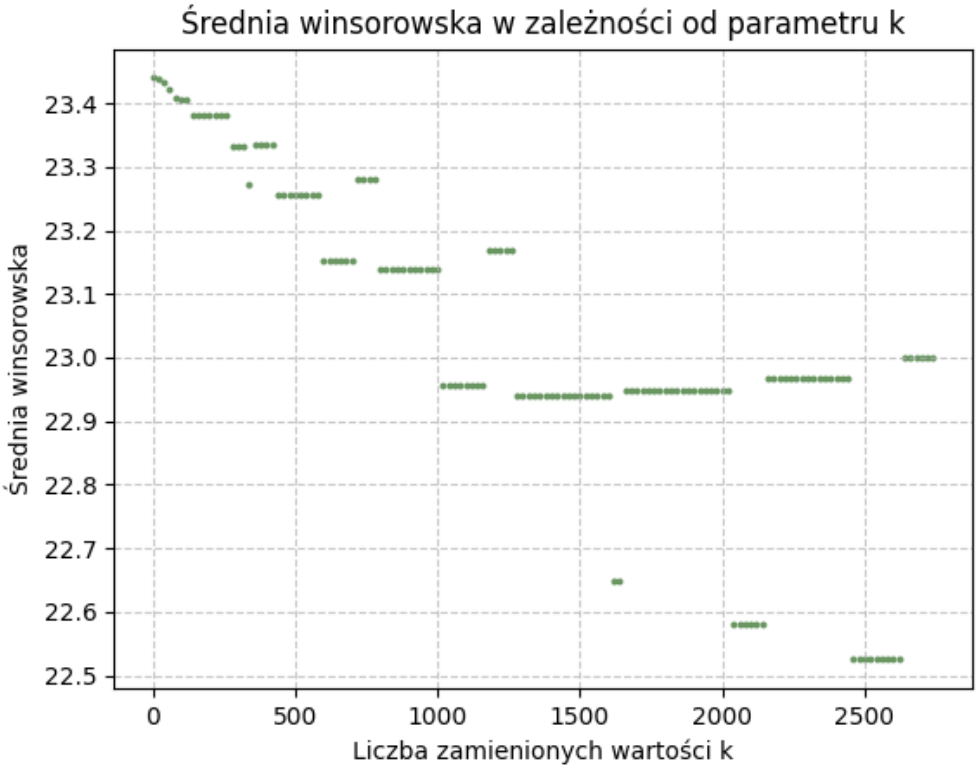
3.5 Średnia ucinana



Rysunek 5: Wykres średniej ucinanej

Po odrzuceniu ok. 2500 ekstremalnych obserwacji można zauważyć, że średnia wieku zmalała aż o ponad $3,5$ (lata). Średnia ucinana jest malejąca, co potwierdza prawostronną skośność badanej próby.

3.6 Średnia winsorowska



Rysunek 6: Wykres średniej winsorowskiej

Spadek średniej winsorowskiej po zmianie ok. 2500 obserwacji z obu stron jest dużo mniejszy niż w przypadku średniej ucinanej, ponieważ wynosi zaledwie ostatecznie ok. $0,5$ (roku).

4 Podsumowanie

Badanym zestawem danych był wiek zawarcia pierwszego małżeństwa kobiet mieszkających w USA, które odpowiedziały na National Survey of Family Growth (NSFG) przeprowadzone przez Centers for Desease Control and Prevention (CDC) w cyklu lat 2006-2010.

W pierwszej kolejności obliczono statystyki opisowe, takie jak średnia, mediana, odchylenie standardowe itp., które pozwalają na dokładne określenie charakterystyki rozkładu danych których wyniki analizy zostały przedstawione w tabelach. Dane również zostały poddane analizie graficznej. Wykonano histogramy, które pokazują, jak często występują różne wartości w danych, a także wykresy gęstości, które pozwalają na określenie kształtu rozkładu. Przy użyciu dystrybuant oraz kwantylowych wykresów można było dokładniej zbadać wartości występujące w danych i określić, jakie jest ich rozmieszczenie wokół średniej wartości. Przy użyciu wykresu kwantylowego zauważono, że odbiega on od wyresu kwantylowego rozkładu normalnego. Wykres pudełkowy pozwolił odczytać niekóre obliczone wcześniej wartości.

Po przeprowadzonej analizie można stwierdzić, że dane nie reprezentują rozkładu normalnego. Wykresy jak i tabele jednodnoznacznie pokazują, że mimo iż na pierwszy rzut oka możnaby pomyśleć, że badane dane pochodzą z wspomnianego wyżej rozkładu normalnego, to po dogłębnym sprawdzeniu tak nie jest.