MS sem.IV **Raport 1 ze Statystyki stosowanej** 2024/04/24

Analiza danych statystycznych Igrzysk Olimpijskich kobiet i mężczyzn: od Aten 1896 do Rio 2016

Zofia Polak, Roksana Rymarek

Statystyka stosowana

Spis treści

1 Wstęp

1.1 Podstawy teoretyczne

Statystyka stanowi nieodłączną część naszego życia, oferując narzędzia do zrozumienia i analizy zjawisk występujących w różnych dziedzinach. Jej istotą jest zbieranie, organizowanie oraz interpretacja danych dotyczących różnorodnych zjawisk lub procesów. Poprzez wykorzystanie metod statystycznych możemy lepiej zrozumieć świat, wyciągać wnioski oraz podejmować decyzje oparte na faktach.

Podstawowe pojęcia statystyczne obejmują różne elementy zbioru danych takie jak:

- Zbiorowość statystyczna to grupa danych, które mają pewną wspólną cechę lub charakterystykę,
- Próba to mniejszy fragment danych, który jest analizowany, aby wyciągnąć wnioski na temat całej zbiorowości,
- Cechy statystyczne są właściwościami obiektów w zbiorowości, które mogą być zmierzone lub opisane,
- Rozmiar próby odnosi się do liczby obserwacji, które zostały zebrane w badaniu.

1.2 Cel pracy

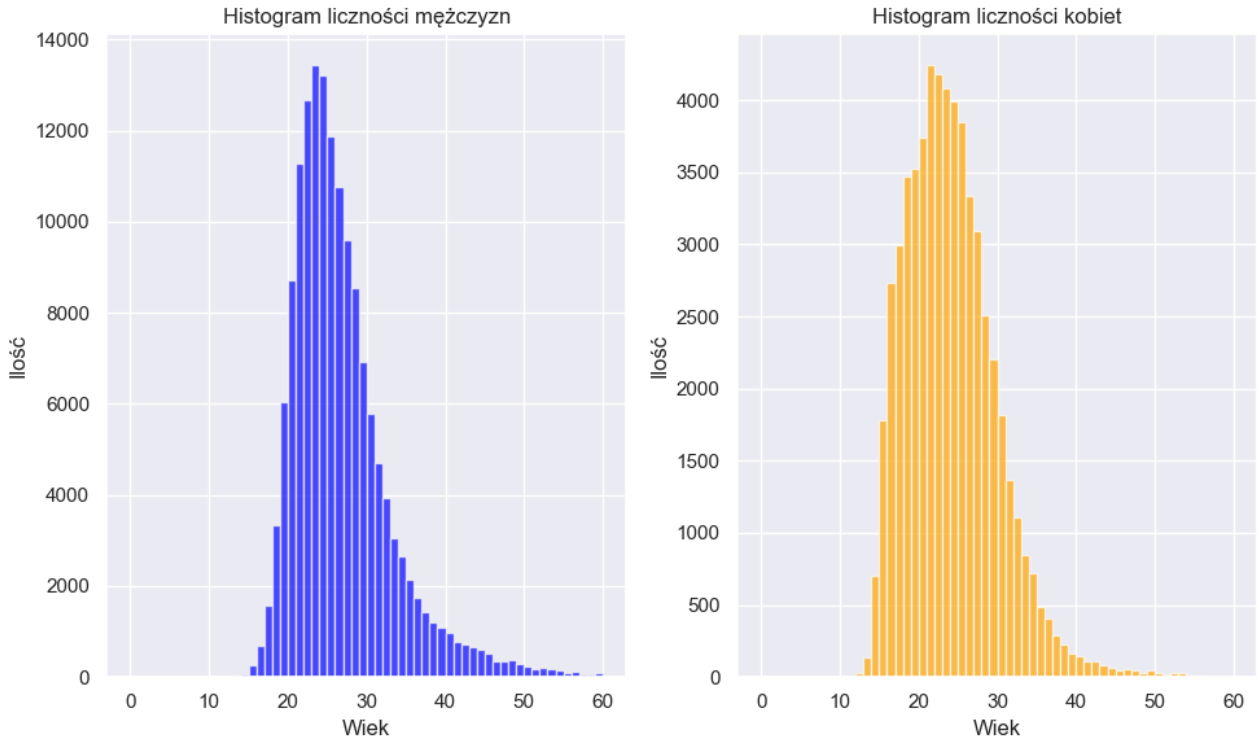
Celem niniejszego badania jest przeprowadzenie analizy danych dotyczących letnich Igrzysk Olimpijskich od Aten 1896 do Rio 2016 w celu identyfikacji kluczowych wykonawców w poszczególnych kategoriach sportowych oraz zbadania czynników determinujących lub różnicujących wyniki medalowe zdobywanych przez sportowców. Poprzez analizę danych historycznych chcemy lepiej zrozumieć dynamikę i ewolucję sportu olimpijskiego, oraz zidentyfikować wzorce sukcesu i wyzwania, jakie stają przed sportowcami na przestrzeni lat. Korzystamy ze zbioru danych, który dotyczy wydarzeń olimpijskich. Dostępne szczegółowe informacje to imię i nazwisko zawodnika, wiek, wzrost, waga, płeć, kategoria sportowa, sezon i medale zdobyte w danej kategorii. Zbiór danych zdarzeń zawiera około 271116 rekordów z 15 zmiennymi. Wszystkie obliczenia i analizy wykonywane były w języku Python.

2 Wizualizacja danych

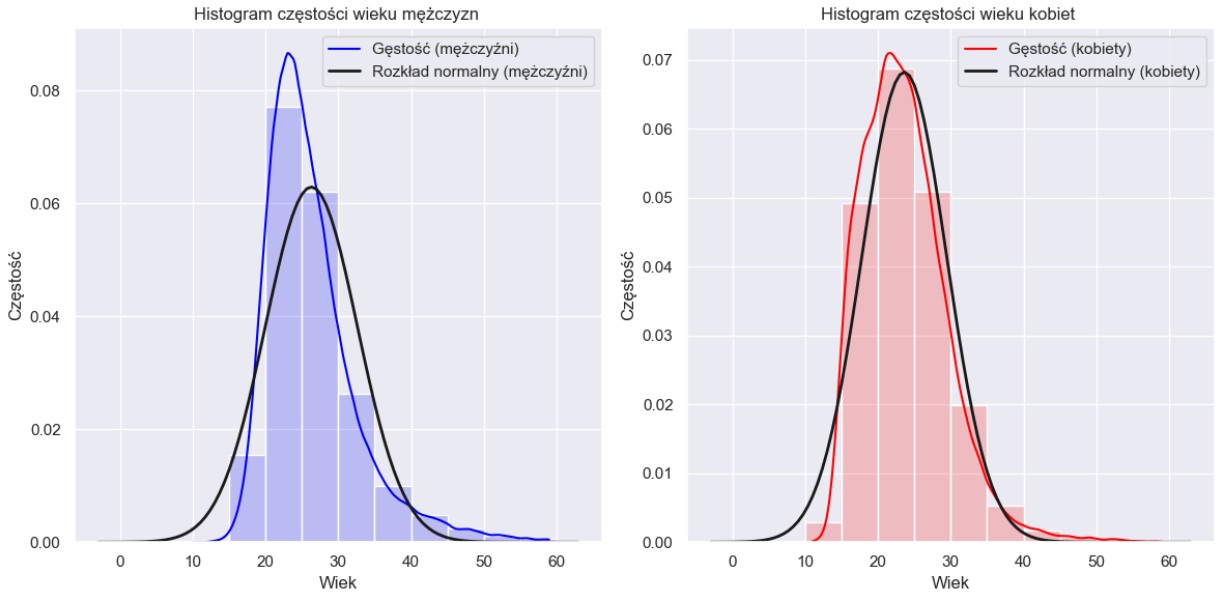
Aby móc lepiej przyjrzeć się danym, przedstawiono je graficznie za pomocą wykresów o różnej charakterystyce.

2.1 Histogramy

Jednym ze sposobów przedstawienia danych jest używanie histogramu liczności, który prezentuje, jak często występuje konkretny wzorzec w analizowanej próbie. Aby móc porównać dwa zestawy danych, wykonuje się histogramy częstości, które są w rzeczywistości empirycznymi odpowiednikami gęstości rozkładu.



Naniesiono również wykresy gęstości empirycznych oraz gęstości rozkładu normalnego, aby sprawdzić, czy pokrywają się z gęstościami naszych prób.



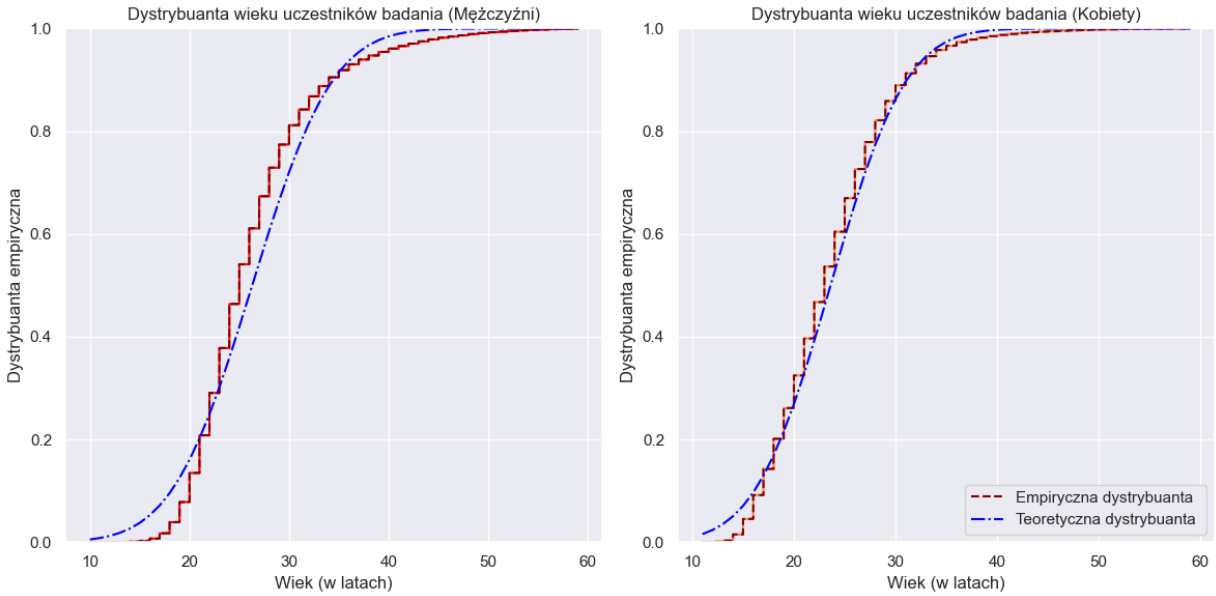
Na podstawie histogramu dla mężczyzn można zauważyć, że największa częstość wieku zawodników znajduje się w przedziale, który prawdopodobnie odpowiada wiekowi około 20-30 lat, co sugeruje, że średni wiek zawodników pozostaje stabilny i skoncentrowany wokół tej dekady życia. Nie obserwujemy znaczących zmian w rozkładzie wieku zawodników męskich w analizowanym okresie, co wskazuje na stałość warunków fizycznych wymaganych do uczestnictwa w sportach olimpijskich.

Histogram dla kobiet również ukazuje modę w podobnym przedziale wiekowym co u mężczyzn, co może sugerować podobieństwo w wieku szczytowych osiągnięć sportowych pomiędzy płciami. Podobnie jak w przypadku mężczyzn, rozkład wieku sportowców kobiet nie wykazuje drastycznych zmian, utrzymując się na stabilnym poziomie.

Porównując gęstości rozkładu normalnego i empiryczne dla obu płci, można dostrzec, że histogramy nie pokrywają się całkowicie z krzywymi rozkładu normalnego, co może świadczyć o tym, że dane nie wykazują idealnie normalnego rozkładu wieku. Jest to widoczne na obu histogramach, gdzie empiryczne gęstości nieco odstają od teoretycznych krzywych rozkładu normalnego.

2.2 Dystrybuanty

Kolejnym sposobem na graficzne zilustrowanie danych, jest wykonanie wykresów dystrybuanty empirycznej, której wartość w punkcie x jest równa częstości zdarzenia polegającego na tym, że obserwacje w próbce są mniejsze od wartości x. Został naniesiony również wykres dystrybuanty rozkładu normalnego w celu ponownej weryfikacji zależności między danymi.

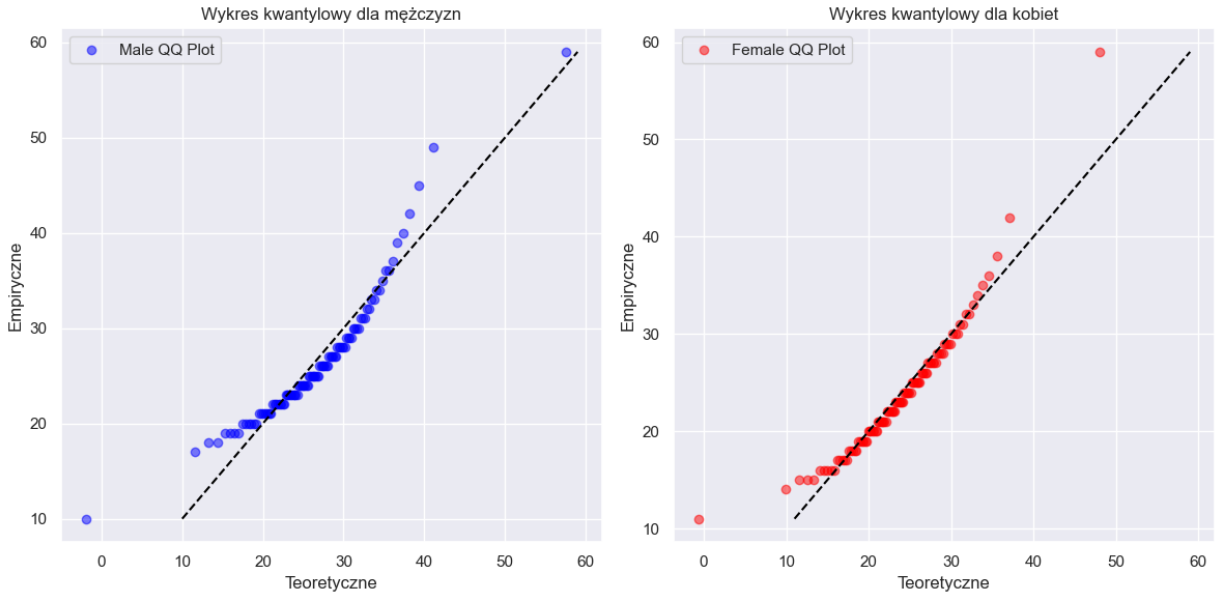


Obie dystrybuanty pokazują dość gwałtowny wzrost w okolicach wieku 20 lat, co wskazuje na to, że wiele osób zaczyna uczestnictwo w Igrzyskach Olimpijskich już w młodym wieku. To potwierdza obserwacje z histogramów, gdzie największa gęstość wieku uczestników skupiona była wokół 20–30 lat.

Obie dystrybuanty są podobne w kształcie, co świadczy o podobnej strukturze wiekowej zarówno wśród mężczyzn, jak i kobiet. Oznacza to, że kobiety i mężczyźni biorący udział w Igrzyskach Olimpijskich mają podobny rozkład wieku.

2.3 Wykresy kwantylowe

Aby sprawdzić, czy dane faktycznie pasują do rozkładu normalnego, można skorzystać z wykresów kwantylowych. Ten typ wykresu umożliwia wizualną analizę, porównując dane empiryczne z teoretycznym rozkładem. Dzięki nim możemy lepiej zrozumieć strukturę danych i zauważyć ewentualne różnice w stosunku do rozkładu normalnego.



Większość punktów na wykresie mężczyzn leży blisko linii odniesienia (przerywanej), co sugeruje, że dane są w przybliżeniu normalnie rozłożone. Istnieją pewne odchylenia, szczególnie w dolnej części wykresu (niższe kwantyle), gdzie punkty oddalają się od linii odniesienia. Wskazuje to na lekką skośność lub większą zmienność w niższych wartościach danych. Kilka punktów w górnej części wykresu również odbiega od linii odniesienia, co może wskazywać na obecność wartości odstających.
Na wykresie kobiet, tak jak w przypadku mężczyzn, większość danych jest blisko linii odniesienia, co może świadczyć o normalności rozkładu. Odchylenia są mniejsze niż w przypadku mężczyzn, co wskazuje na bardziej symetryczny rozkład wokół średniej. Mimo to, pojawiają się pewne odchylenia, zwłaszcza w ekstremalnych kwantylach, sugerujące obecność wartości odstających lub dłuższych ogonów rozkładu.

3 Statystyki opisowe

Analiza danych rzeczywistych wymaga wykorzystania metod i statystyk opisowych, które pozwalają na zrozumienie charakterystyki badanej próby oraz podsumowanie zebranych informacji w celu wyciągnięcia odpowiednich wniosków. Istotnym narzędziem w tym procesie jest określanie miar rozkładu, które dostarczają informacji na temat struktury badanego rozkładu. W dalszej części pracy, litera "n" będzie używana do oznaczenia rozmiaru próby.

3.1 Miary położenia

Miary położenia są używane do identyfikowania typowych wartości obecnych w analizowanej próbie. Istnieją dwie główne kategorie tych miar: klasyczne, które obejmują miary takie jak średnie arytmetyczne oraz pozycyjne, które obejmują miary takie jak kwartyle. Zarówno miary klasyczne, jak i pozycyjne, pozwalają na zrozumienie różnych aspektów charakterystyki badanej próby.

Podstawowe rodzaje średnich:

Średnia arytmetyczna: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Średnia harmoniczna: $H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

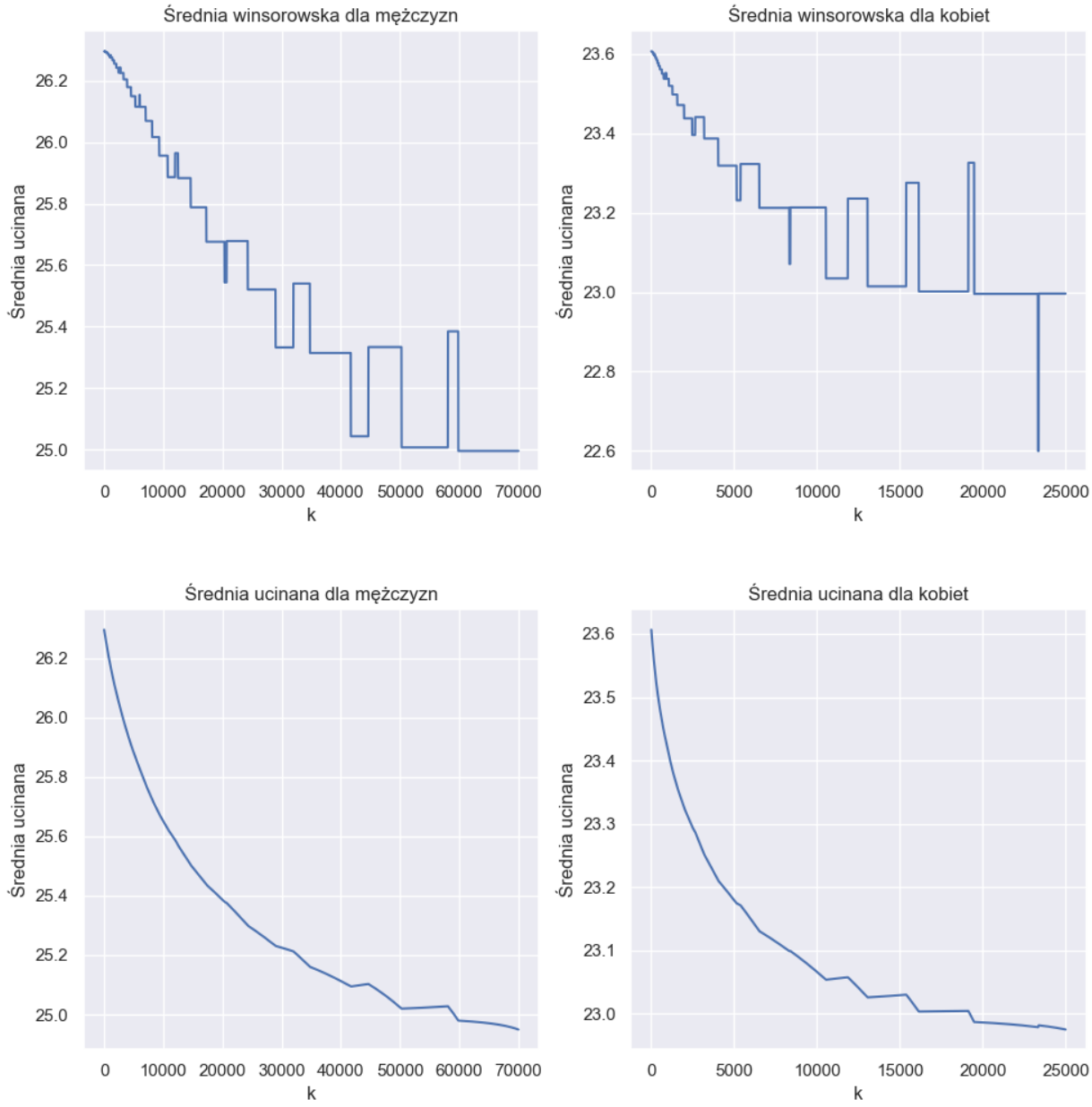
Średnia geometryczna: $\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Średnia winsorowska: $\bar{x}_w = \frac{1}{n}\left((k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}\right)$

Średnia ucinana: $\bar{x}_u = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)}$

W tabeli przedstawiono wartości powyższych miar dla badanych zestawów danych.

	Wiek kobiet	Wiek mężczyzn
Średnia arytmetyczna	23.61	26.30
Średnia harmoniczna	22.30	25.05
Średnia geometryczna	22.94	25.64
Średnia winsorowska	23.00	25.00
Średnia ucinana	23.00	25.00



Zarówno dla mężczyzn, jak i kobiet, szybkie zmniejszanie się wartości średniej ucinanej przy niewielkich wartościach k i jej późniejsze ustabilizowanie wskazują na brak znacznej liczby ekstremów w danych. Średnia ucinana nieznacznie rośnie po usunięciu najniższych wartości, co może świadczyć o lekkiej lewoskośności rozkładu.

Na wykresach średniej winsorowskiej obserwujemy większe wariacje przy większych wartościach k, co sugeruje obecność ekstremów. W przypadku mężczyzn widać duże skoki, wskazujące na bardziej zgrupowane ekstremalne wartości. Dla obu płci, przy dużych

wartościach k, średnia winsorowska zbliża się do mediany, co jest zgodne z oczekiwaniami winsorowania.

Kwartyle to statystyki pozycyjne dzielące uporządkowany zbiór danych na cztery równe części. Kluczowe kwartyle to:

- **Mediana (Q2)**: Znana również jako drugi kwartyl, dzieli zbiór danych na dwie połowy. Dla zbioru o nieparzystej liczbie elementów n, mediana jest wartością środkową $x_{\frac{n+1}{2}}$. Gdy liczba elementów jest parzysta, mediana jest średnią dwóch środkowych wartości, czyli $\frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$.
- **Pierwszy kwartyl (Q1)**: Jest to mediana dolnej połowy danych, nie uwzględniająca mediany (Q2), czyli środkowej wartości zbioru danych mniejszych od mediany.
- **Trzeci kwartyl (Q3)**: Stanowi mediana górnej połowy danych, wyznaczona analogicznie do pierwszego kwartyla, dla wartości wyższych od mediany.

Wartości kwartyli dla analizowanej serii danych zostały zaprezentowane w poniższej tabeli:

	Wiek mężczyzn	Wiek kobiet
Q1	22.0	19.0
Mediana	25.0	23.0
Q2	29.0	27.0

3.2 Miary rozproszenia

Miary rozproszenia dostarczają informacji o stopniu zróżnicowania danych w zbiorze. Do głównych miar rozproszenia należą:

Rozstęp międzykwartylowy (IQR): $IQR = Q3 - Q1$

Rozstęp z próby (R): $R = x_{(n)} - x_{(1)}$

Wariancja (S^2): $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Odchylenie standardowe (S): $S = \sqrt{S^2}$

Współczynnik zmienności (V): $$V = \frac{S}{\bar{x}} \times 100\%$$

Wartości tych miar dla badań danych zostały przedstawione w poniższej tabeli:

	Wiek mężczyzn	Wiek kobiet
IQR	7.0	8.0
R	49	48
S^2	40.21	34.20
S	6.34	5.85
V	24%	24%

3.3 Inne miary

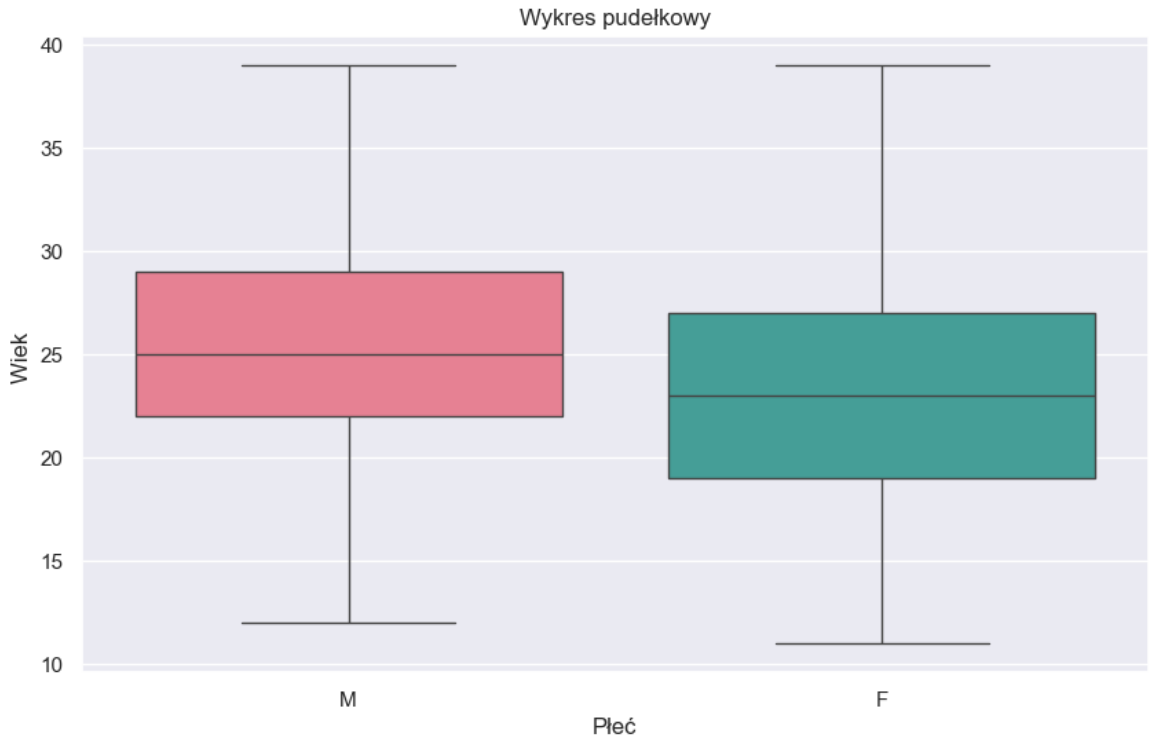
Współczynnik skośności (miara asymetrii): $$\alpha = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^3}{S^3}$$

Kurtoza (miara spłaszczenia): $$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

Wyliczone wartości powyższych dwóch miar przedstawiono w tabeli:

	Wiek mężczyzn	Wiek kobiet
Alfa	1.51	1.00
K	3.31	2.11

3.4 Wykresy pudełkowe



Analiza wykresów pudełkowych pozwala na wyciągnięcie różnych wniosków dotyczących wieku mężczyzn i kobiet na Igrzyskach Olimpijskich. Pomimo minimalistycznego wyglądu, wykres ramka-wąsy jest bardzo informacyjny, przedstawiając tylko pięć podstawowych danych. Wykres pudełkowy składa się z prostokąta (pudełka), osi współrzędnych oraz tzw. wąsów, które reprezentują odległość od minimalnej do maksymalnej wartości w danych. Na podstawie naszego wykresu możemy zauważyć, że maksymalna wartość wieku mężczyzn i kobiet to około 40 lat, natomiast minimalna to powyżej 10 lat. Linia pozioma wewnątrz prostokąta na wykresie pudełkowym reprezentuje medianę, czyli wartość środkową próbek.W naszym przypadku wynoszą one kolejno 25 lat i około 23 lata. Mediana mężczyzn leży wyraźnie poniżej średnią, co potwierdza, że rozkład tych danych jest prawostronnie skośny. W drugim zestawie danych nie jest to zbyt zauważalne.

4 Podsumowanie

Zestaw danych obejmuje historię nowożytnych igrzysk olimpijskich od Aten 1896 do Rio 2016. W szczególności, skupiliśmy się na wieku sportowców, zarówno mężczyzn,

jak i kobiet w sezonie letnim. Dane zostały poddane analizie graficznej, w ramach której wykonano histogramy, wykresy gęstości, dystrybuant oraz wykresy kwantylowe. Po tej wstępnej wizualizacji przeprowadzono obliczenia znanych statystyk opisowych, takich jak średnia, mediana itp. Wyniki tych obliczeń zostały przedstawione w tabelach, co umożliwiło lepsze zrozumienie charakterystyki danych oraz ich rozkładu. Te kroki są często początkowym etapem analizy danych, pomagającym zidentyfikować ważne tendencje i zależności w zestawie danych. Wszystkie te czynności pozwoliły stwierdzić, że w ciągu ponad 120 lat na igrzyskach najczęściej brały udział mężczyźni w wieku 25 lat oraz kobiety w wieku 23 lat.

5 Literatura

Nasze dane : 120 years of Olympic history: athletes and results

https://www.kaggle.com/code/josephgpinto/holding-an-olympic-games-means-evoking-history/report

Teoria:

https://eks.stat.gov.pl/materialy/scenariusze/miary$_s$*tatystyczne*/*materialy*$_d$*la*$_n$*auczyciela.pdf*