Statystyka

Analiza wzrostu i wagi sportowców

Prowadzący: Katarzyna Maraj-Zygmąt
Autorzy: Maria Nowacka i Bartłomiej Mielcarz

8 maja

1 Cel ćwiczenia

Celem tego raportu jest wykorzystanie poznanych narzędzi statystycznych do analizy dwóch zmiennych losowych: wzrostu i wagi sportowców (kobiet), aby zbadać czy istnieje między nimi korelacja, oraz żeby sprawdzić jaki rozkład mają te zmienne.

2 Źródło danych

Dane użyte w tej analizie pochodzą z:
www.kaggle.com/datasets/mysarahmadbhat/120-years-of-olympic-history?resource=download
Zostały zebrane w okresie 1920 - 2016 (większość danych z lat 1950 - 2016) i opisują pomiary wzrostu (X) i wagi (Y) dla 10 350 sportowców płci żeńskiej.
Symulacje potrzebne do analizy zostały wykonane w języku python.

3 Analiza danych

3.1 Statystyki opisowe

3.1.1 Średnia

Średnia to sposób na określenie typowej wartości w grupie danych, jak na przykład wzrost czy waga sportowców. Odpowiedni typ średniej pozwala lepiej zrozumieć rozkład i tendencje w grupie.

arytmetyczna	170.53
harmoniczna	170.01
geometryczna	170.30
mediana	170.00

Table 1: średnie wartości - wzrost

arytmetyczna	63.22
harmoniczna	61.47
geometryczna	62.35
mediana	63.00

Table 2: średnie wartości - wagi

Inne średnie, ucinana i winsorowska, powinny zbiegać do mediany. Na wykresach możemy zauważyć, że istotnie tak się dzieje przy ucięciu odpowiedniej ilości danych.

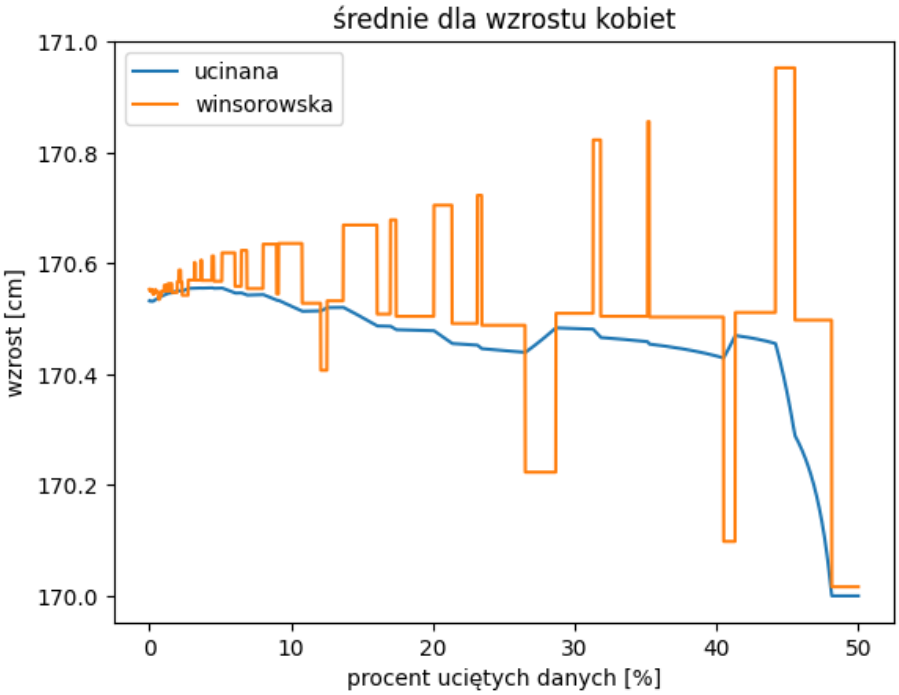

Figure 1: Średnia ucinana i winsorowska dla wzrostu

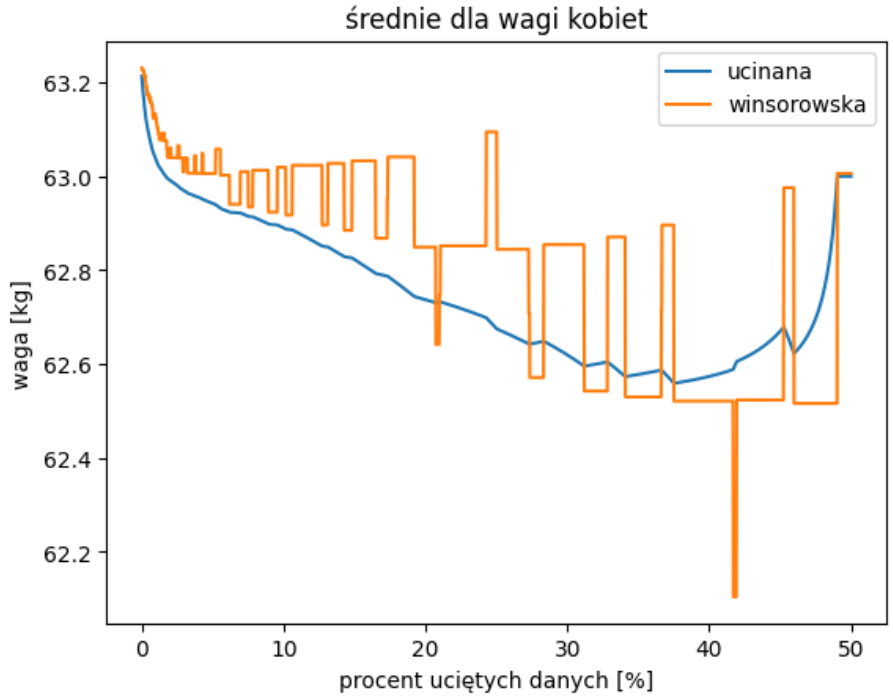

Figure 2: Średnia ucinana i winsorowska dla wagi

3.1.2 Mediana

Mediana to wartość środkowa, dzieląca zbiór na dwie równe części. Gdy analizujemy na przykład wzrost sportowców, mediana pokazuje wzrost środkowego sportowca.

3.1.3 Odchylenie standardowe

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

Odchylenie standardowe informuje nas o rozproszeniu danych wokół średniej. Im większe odchylenie standardowe, tym mniejsza jednorodność. W naszym przypadku odchylenie standardowe wynosi **8.98** dla wzrostu, a dla wagi **10.68**.

Takie odchylenie standardowe dla wzrostu, przy średniej arytmetycznej wynoszącej około 170 można uznać za w miare duże. Z tego wynika, że występuje duże zróżnicowanie wzrostu. Większość sportowców ma wzrost bliski okolicy średniej, ale istnieje spora liczba osób o wzroście znacznie niższym lub wyższym. W przybliżeniu można założyć, że około 68,3% wszystkich sportowców ma wzrost w przedziale od 161 cm do 179 cm. Około 95,5% sportowców będzie miało wzrost w przedziale od 152 cm do 190 cm.

Obliczając empirycznie, jaki procent kobiet mieści się w podanych przedziałach wzrostu uzyskujemy: 68,9% w przedziale 161-179 cm oraz 95% w przedziale 152-190 cm. widzimy, że wyniki te są zbliżone do oczekiwanych.

Odchylenie standardowe około 11 kg przy średniej 63 kg jest stosunkowo większe niż w przypadku wzrostu. To świadczy o tym, że waga sportowców jest bardziej zróżnicowana niż ich wzrost. Około 68,3% sportowców waży między 52 kg a 74 kg, a około 95,5% sportowców waży między 41 kg a 85 kg, co pokazuje szeroki zakres wagowy. Obliczając empirycznie jaki procent kobiet mieści się w podanych przedziałach wagowych uzyskujemy: 72,6% w przedziale 52-74 kg oraz 95,8% w przedziale 41-85 kg. Ponownie nasze rezultaty są zbliżone do oczekiwanych wartości dla rozkładu normalnego o odpowiednich parametrach.

Większe odchylenie standardowe w kontekście wagi może wskazywać na większe zróżnicowanie w grupie sportowców, co może być wynikiem różnic w typach ciał lub specyfikach różnych dyscyplin sportowych. Na przykład, sportowcy uprawiający dyscypliny wymagające większej masy ciała lub siły mogą ważyć znacznie więcej. To, że wartości skrajne są bardziej powszechne w przypadku wagi niż wzrostu sugeruje, że waga jest cechą zmienną i jest bardziej podatna na takie czynniki jak dieta czy trening.

4 Wizualizacje danych

4.1 Histogramy

Histogramy dla wzrostu X i wagi Y, pokazujące rozkład danych. Poniższe histogramy i ich gęstości wskazują na fakt, że wartości skrajne są znacznie bardziej widoczne w przypadku wagi, niż w przypadku wzrostu. Możemy też zauważyć, że badane zmienne mają rozkład mocno zblizony do normalnego o parametrach $\mu = 170$ i $\sigma = 8.98$ dla wzrostu oraz $\mu = 63$ i $\sigma = 10.68$ dla wagi.

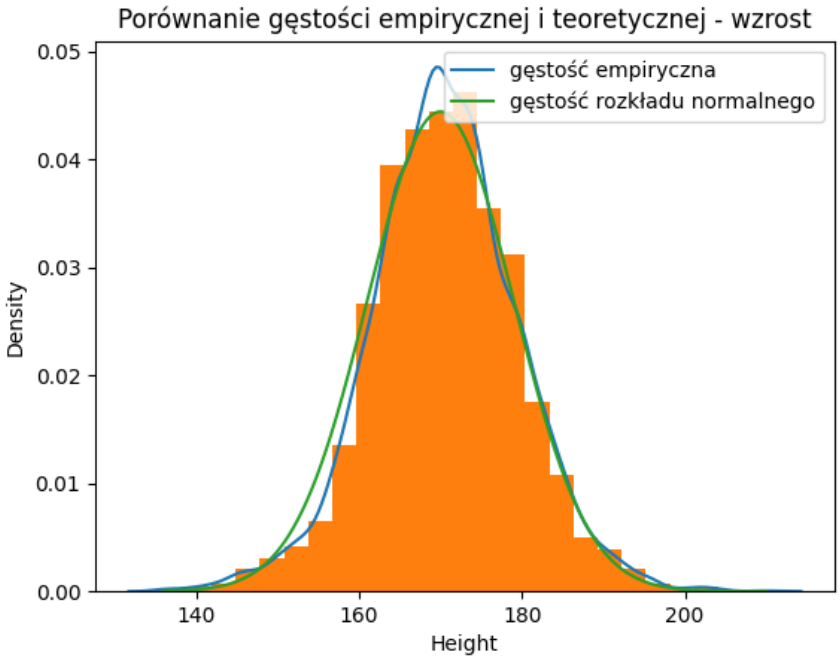

Figure 3: Histogram, gęstość empiryczna i teoretyczna wzrostu

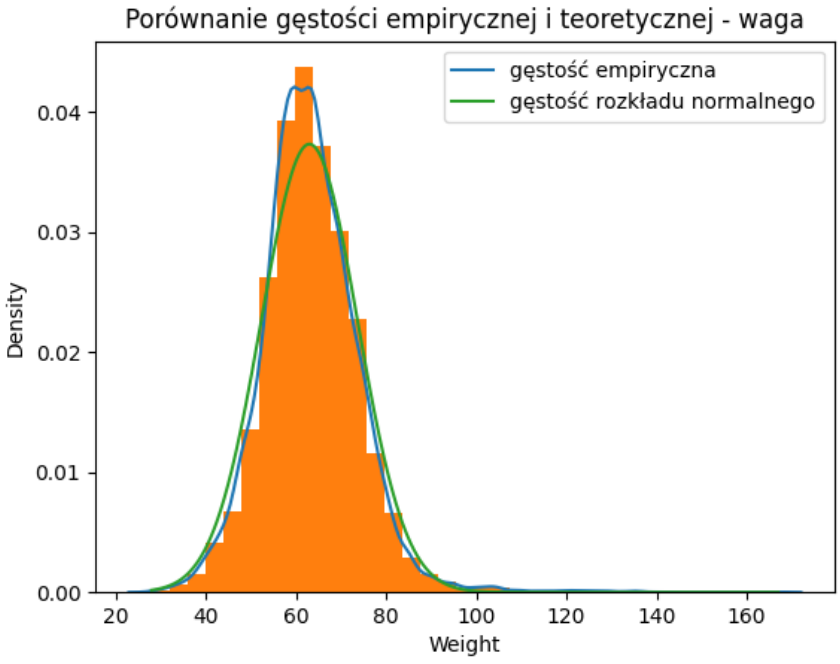

Figure 4: Histogram, gęstość empiryczna i teoretyczna wagi

4.2 Dystrybuanty

Dystrybuanty dla obu zmiennych oraz ich wartości teoretyczne, tzn. z rozkładu normalnego o odpowiednich paametrach.

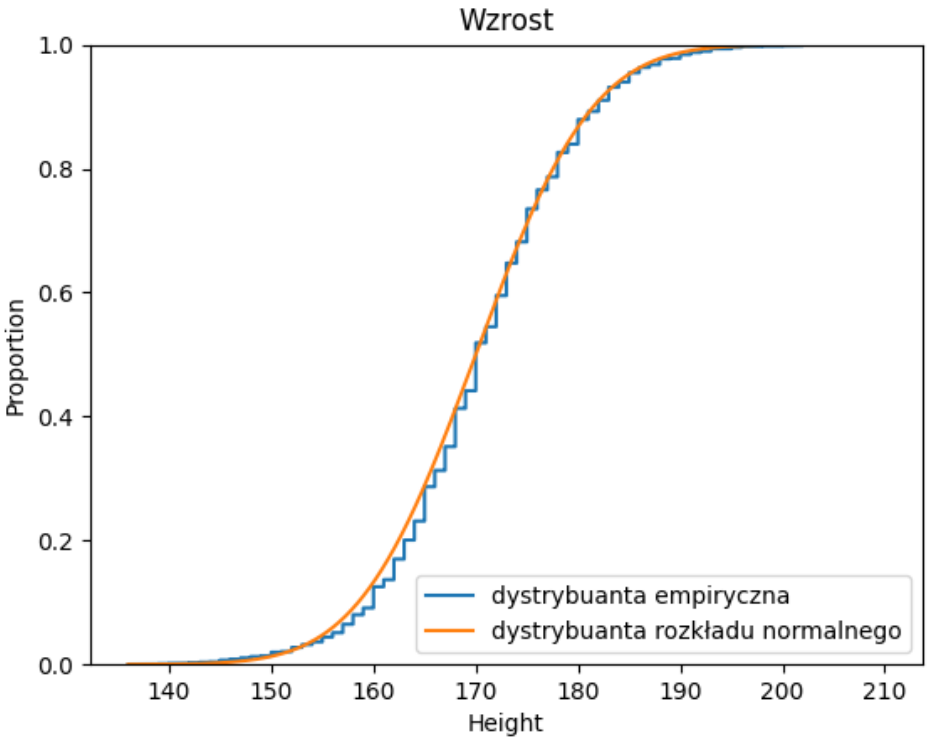

Figure 5: Dystrybuanta empiryczna i teoretyczna wzrostu

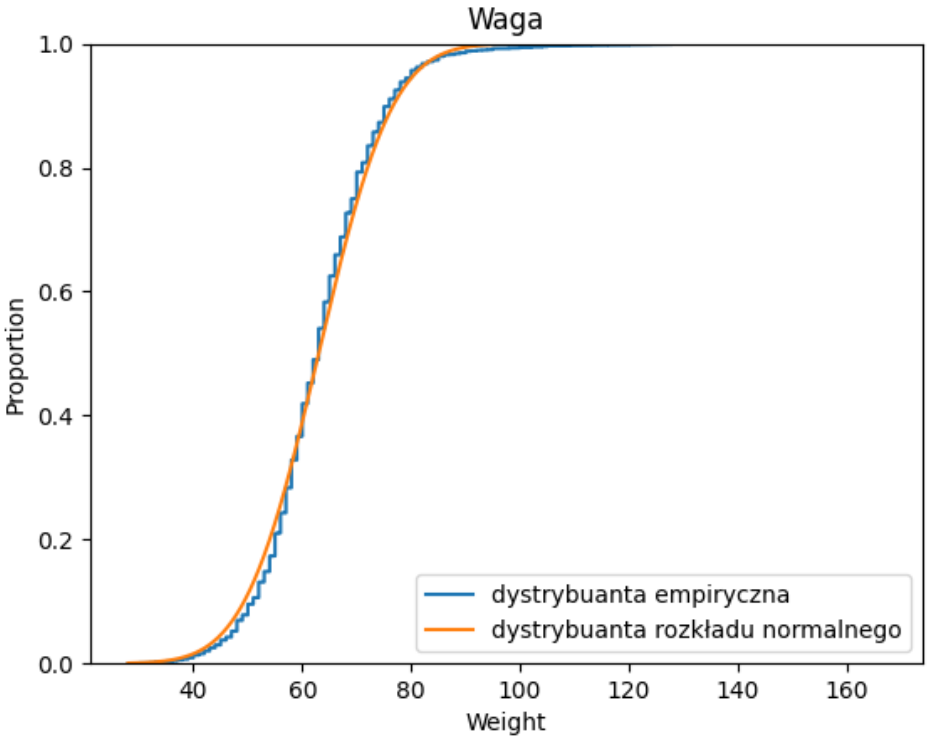

Figure 6: Dystrybuanta empiryczna i teoretyczna wagi

5 Porównanie zmiennych X i Y

5.1 Boxploty

Boxploty dla X i Y, przedstawiające rozkład danych i wartości odstające, z których również można wywnioskować, że odchylenie standardowe wagi jest większe niż odchylenie standardowe wzrostu.

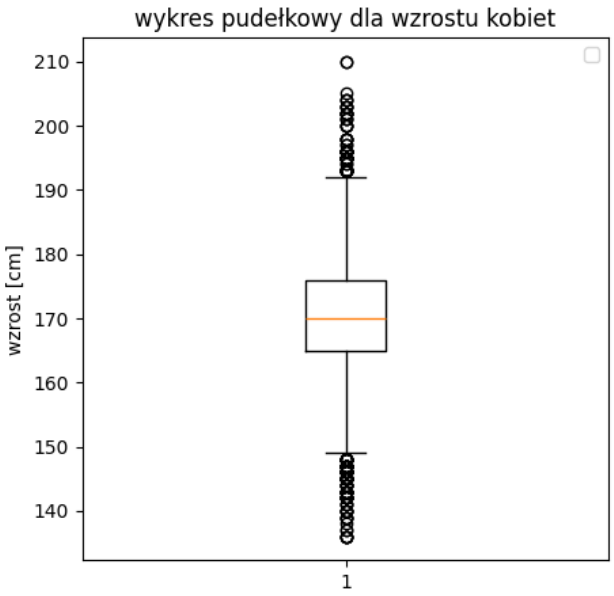

Figure 7: Wykres pudełkowy - wzrost

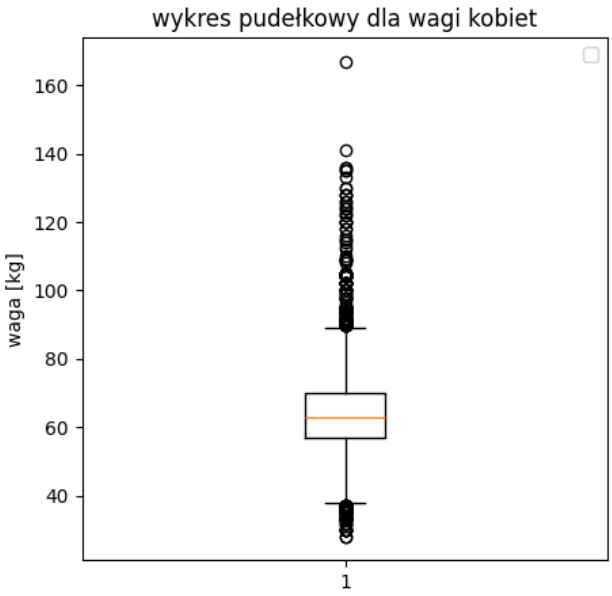

Figure 8: Wykres pudełkowy - waga

5.2 Macierz korelacji

Aby zbadać korelację między danymi, możemy zacząć od sprawdzenia, jak prezenują się one na wykresie i stwierdzić, czy może istnieć między nimi jakaś zależność. Jak widać na wykresie Figure 9, między zmiennymi istnieje pewna korealcja. Dokładną wartość możemy obliczyć używając biblioteki nupmy w języku python. Dla naszych danych macierz korelacji prezentuje się w następujący sposób.

$$\begin{bmatrix} 1.0 & 0.73 \\ 0.73 & 1.0 \end{bmatrix}$$

Możemy z tego wnioskować, że wzrost i waga są dosyć silnie skorelowane.

Oprócz tego możemy również obliczyć współczynnik korelacji Spearmana, który dla naszych danych wynosi **0.75**, co również wskazuje na dość silną zależność między zmiennymi.

6 Podsumowanie

Po przeanalizowaniu i zwizualizowaniu danych możemy stwierdzić, że nasze zmienne losowe X i Y (wzrost i waga sportowców) są ze sobą dosyć silnie skorelowane oraz obie mają rozkład

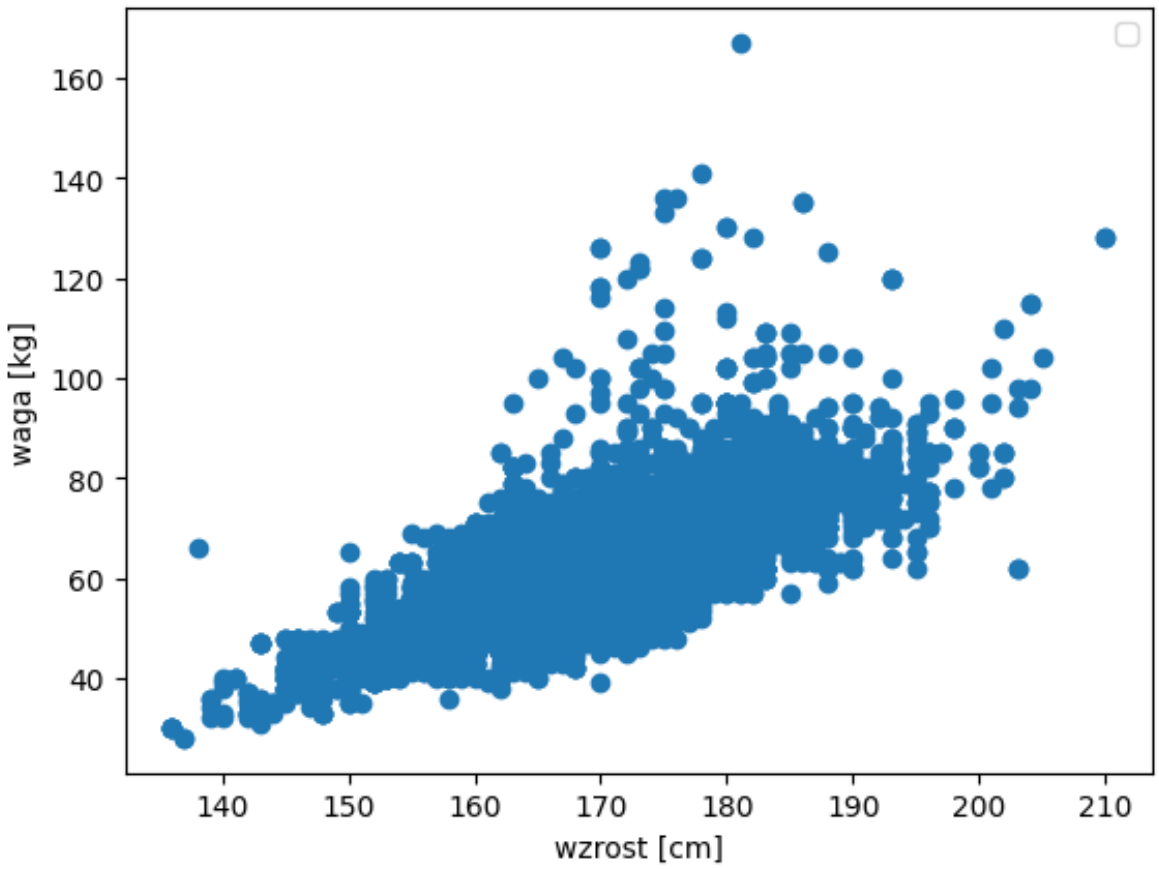

Figure 9: Wykres zależności zmiennych

normalny. Potwierdza to wiedzę z podstaw genetyki - cechy ilościowe (mierzalne) przyjmują wartości będące częścią rozkładu normalnego. Potwierdziliśmy to, generując na wykresach gęstości oraz dystrybuanty teoretyczne oraz empiryczne - linie wykresów nachodzą lub znajdują się bardzo blisko siebie.

Wysoki współczynnik korelacji sugeruje, że waga rośnie wraz ze wzrostem. Jest to dosyć naturalne założenie, ponieważ zazwyczaj osoby wyższe mają wyższą wagę.

Kolejną obserwacją jest większe odchylenie standardowe dla wagi, niż dla wzrostu - może to wynikać z faktu, iż na wagę mamy większy wpływ, a także niektóre sporty wymagają konkretnej wagi u zawodników, przez co zmienna przyjmuje lekko zniekształcone wartości (w porównaniu do teoretycznych wartości z rozkładu normalnego).