Statystyka stosowana

Raport 1

Temat: **Analiza wybranego zbioru danych**
Imię i Nazwisko prowadzącego kurs: **Mgr Katarzyna Maraj-Zygmąt**

Imię i Nazwisko, nr indeksu	Szymon Malec, 262276
Wydział	Wydział matematyki, W13
Dzień i godzina zajęć:	Wtorek, 7^{30}
Kod grupy ćwiczeniowej	T00-64c
Data oddania raportu:	10.05.2022
Ocena końcowa	

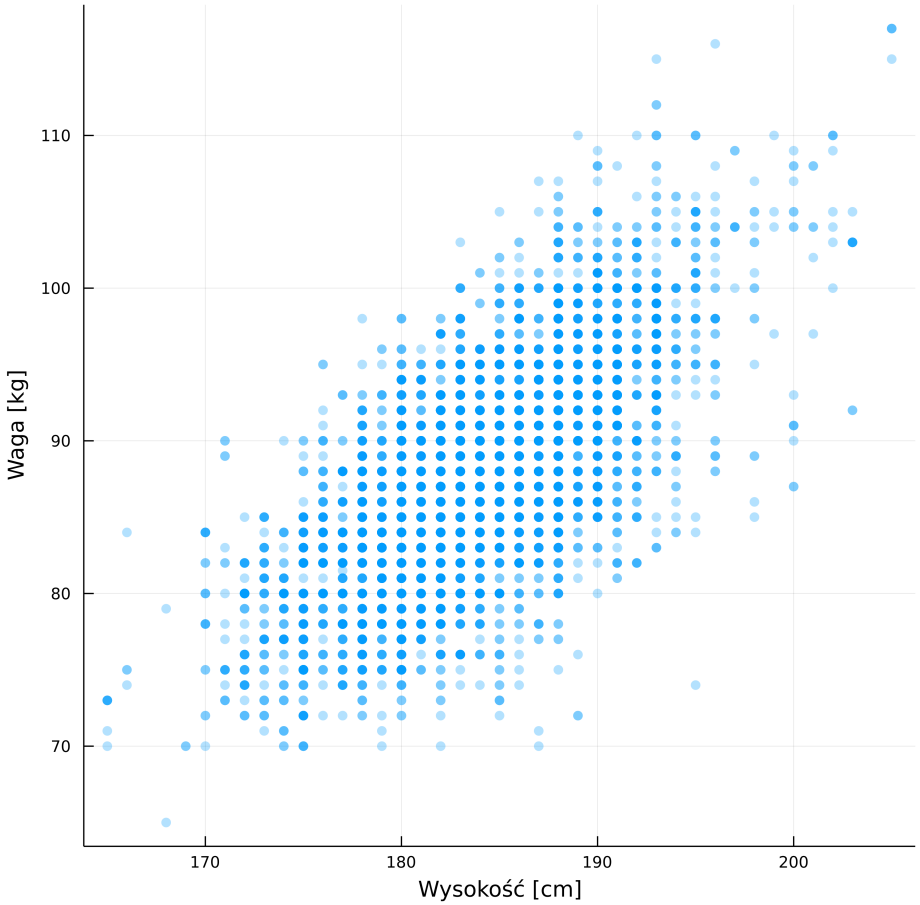
Adnotacje i uwagi:

1. Wstęp

Celem raportu jest wykorzystanie dotychczas poznanych narzędzi statystycznych do zbadania wybranego zbioru danych. W dalszej części przeanalizuję pewną grupę danych w celu wyznaczenia ich podstawowych charakterystyk, znalezienia ich rozkładu oraz sprawdzenia czy dane w jakiś sposób ze sobą korelują.

2. Opis danych[1]

Do opracowania wybrałem zbiór danych zawierający 6292 pomiary wzrostu i wagi 3333 zawodników startujących w mistrzostwach świata w hokeju na lodzie mężczyzn (IIHF World Championship) w latach 2001-2016. Ponieważ dane pochodzą z 16 turniejów, niektórzy zawodnicy, którzy startowali więcej niż jeden raz, powtarzają się. Nie pomijamy jednak tych powtórzeń, ponieważ mają one wpływ na całkowity rozkład wysokości i wagi osób startujących w zawodach.

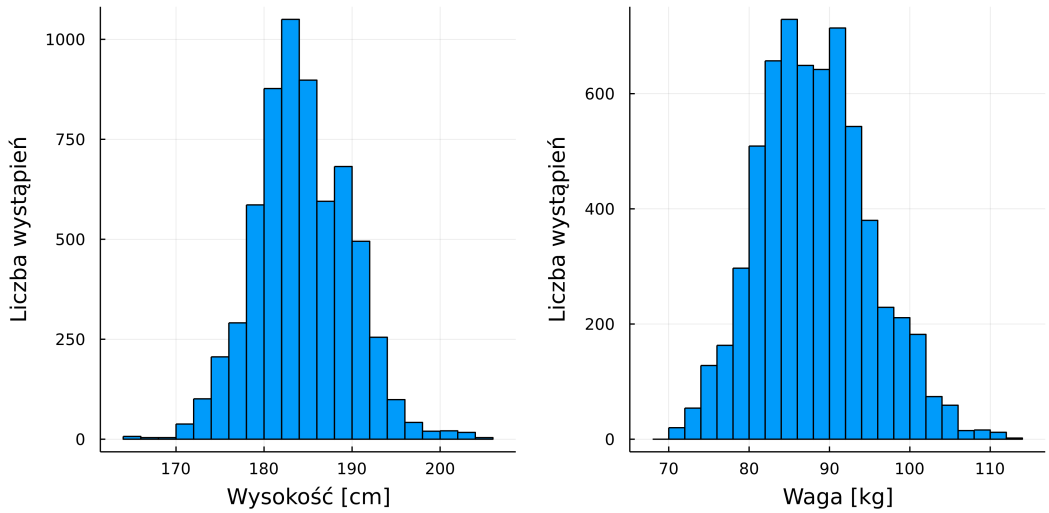


Rysunek 1: Wykres punktowy wysokości od wagi zawodnika.

Wysokość [cm]	**Waga [kg]**
185	84
188	86
182	95
178	85
175	88
193	93
176	84
183	91
180	85
178	86
187	93
185	80
198	95
175	77
178	75
181	85
194	95
186	87
184	86
178	92
175	90
187	86
180	75
188	84
182	88
182	88
175	88
188	90
190	96
181	86
196	98
190	100
185	95
186	88
180	76
176	72
183	82
173	85
185	87
182	84

Wysokość [cm]	**Waga [kg]**
191	94
182	92
184	90
182	92
192	90
187	80
179	88
180	78
178	78
188	84
180	78
178	83
183	91
182	78
185	87
193	97
179	83
185	82
185	83
184	92
186	80
184	89
178	83
178	88
185	87
192	99
190	92
191	92
193	97
183	89
183	82
184	84
177	80
177	77
178	88
178	81
181	84
185	90
181	88
178	86

Tabela 1: Przykładowe dane.



Rysunek 2: Po lewej histogram z wysokości, po prawej z wag.

3. Analiza rozkładów wysokości i wagi zawodników

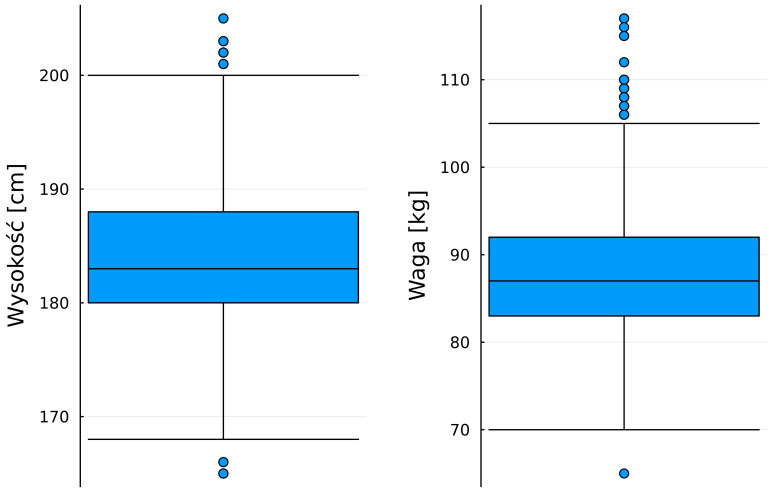
Dane mają charakter dyskretny, co jest spowodowane dokładnością pomiarów wynoszącą 1cm w przypadku wysokości i 1kg w przypadku wagi. Rzeczywiste wartości mają jednak rozkład ciągły. Rozkład dyskretny, który otrzymamy z danych, będzie można traktować jako przybliżenie rzeczywistego rozkładu ciągłego.

Niech $x_1, x_2, \ldots, x_n$ oznaczają dane posortowane rosnąco. Liczba danych wynosi $n = 6292$. Obliczamy podstawowe charakterystyki, w celu otrzymania więcej informacji o rozkładach.

Charakterystyka	**Wzór**	**Wartość dla wysokości**	**Wartość dla wagi**
Średnia arytmetyczna	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	183,8	87,6
Wariancja	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	29	48,5
Odchylenie standardowe	$S = \sqrt{S^2}$	5,4	7
Mediana	——	183	87
Kwartyl Q_1	——	180	83
Kwartyl Q_3	——	188	92
Rozstęp	$R = x_n - x_1$	40	52
Rozstęp międzykwartylowy	$IQR = Q_3 - Q_1$	8	9
Współczynnik zmienności	$V = \frac{S}{\bar{x}}$	0,029	0,079

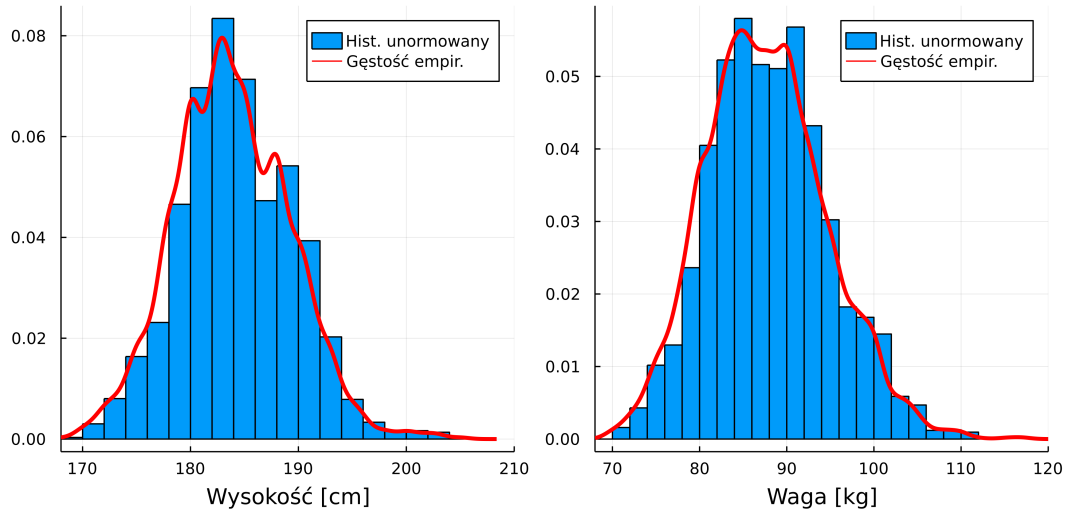
Tabela 2: Najważniejsze charakterystyki obliczone dla danych

Na powyższej tabeli możemy zobaczyć, że w obu przypadkach średnia arytmetyczna jest zbliżona do mediany, co wskazuje, że rozkłady te są symetryczne. Widzimy także, że odchylenie standardowe dla wag jest trochę większe niż dla wysokości, zatem te pierwsze dane mają nieco większy rozrzut. Wskazuje na to także rozstęp międzykwartylowy.



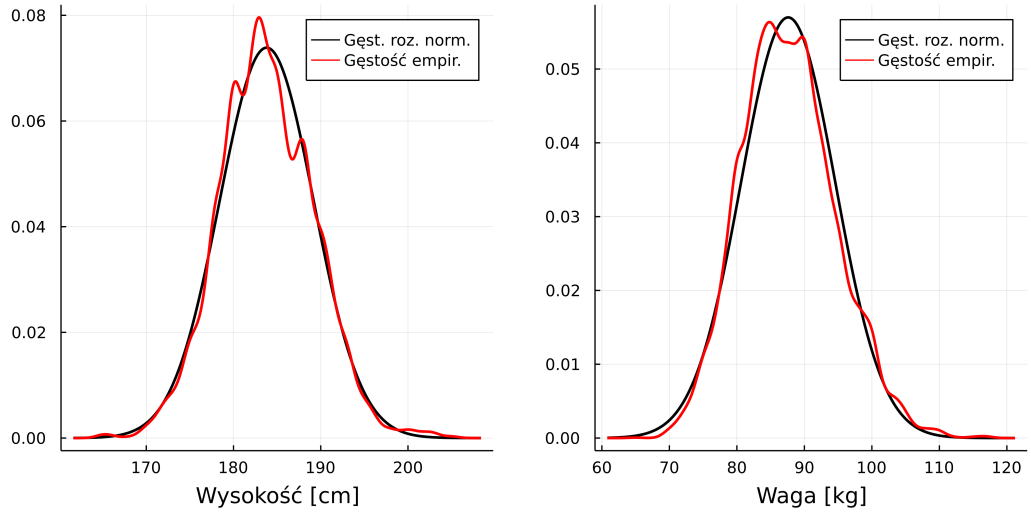
Rysunek 3: Wykresy pudełkowe dla badanych danych.

Korzystając z pakietu KernelDensity dostępnego w Julii możemy znaleźć gęstość empiryczną dla naszych danych, która będzie przybliżeniem gęstości rozkładu, z którego te dane pochodzą.



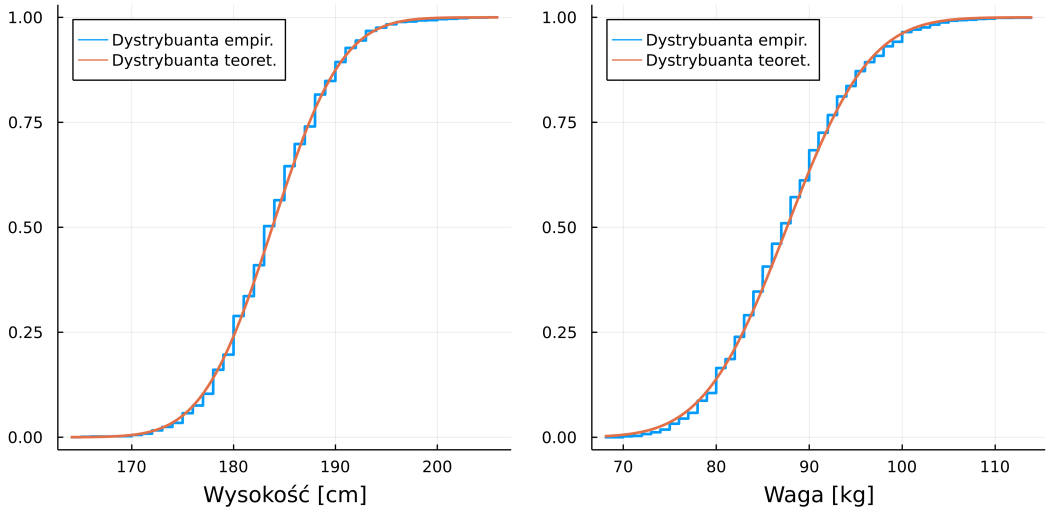
Rysunek 4: Porównanie unormowanych histogramów i gęstości empirycznych, po lewej dla wysokości, po prawej dla wagi.

Jak możemy zobaczyć na powyższych wykresach, gęstości empiryczne pokrywają się z histogramami. Dodatkowo zauważamy, że w obu przypadkach krzywe gęstości przypominają krzywą Gaussa. Aby sprawdzić czy w rzeczywistości nasze dane pochodzą z rozkładu normalnego, porównamy wykresy gęstości empirycznych z wykresami gęstości teoretycznych rozkładu $\mathcal{N}(\mu, \sigma)$, gdzie pod μ podstawimy obliczone wcześniej średnie arytmetyczne, a pod σ odchylenia standardowe.



Rysunek 5: Porównanie gęstości empirycznej otrzymanej z danych z gęstością rozkładu normalnego $\mathcal{N}(\mu, \sigma)$. Po lewej dla wysokości: $\mu = 183{,}8$, $\sigma = 5{,}4$. Po prawej dla wagi: $\mu = 87{,}6$, $\sigma = 7$.

Okazuje się, że krzywe w obu przypadkach wyraźnie się pokrywają. Dodatkowo dla pewności możemy porównać jeszcze wykresy dystrybuanty empirycznej z dystrybuantą teoretyczną rozkładu normalnego.



Rysunek 6: Porównanie dystrybuanty empirycznej otrzymanej z danych z dystrybuantą teoretyczną rozkładu normalnego $\mathcal{N}(\mu, \sigma)$. Po lewej dla wysokości: $\mu = 183{,}8$, $\sigma = 5{,}4$. Po prawej dla wagi: $\mu = 87{,}6$, $\sigma = 7$.

Podobieństwo krzywych widocznych powyżej potwierdza, że dane pochodzą najprawdopodobniej z rozkładu normalnego lub rozkładu bardzo do niego zbliżonego.

4. Analiza korelacji pomiędzy wagą, a wysokością

Jak możemy zauważyć na rys. 1, wartości wydają się być od siebie zależne. Nie jest to zaskoczeniem, ponieważ naturalnym jest, że waga zależy od wysokości człowieka. Policzmy współczynnik korelacji Pearsona, by przekonać się jak mocno te dwa zbiory są ze sobą skorelowane. Niech $x_1, \ldots, x_n$ oznaczają wysokości, a $y_1, \ldots, y_n$ wagi. Wtedy

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x S_y} \approx 0{,}92,$$

gdzie S_x i S_y to odchylenia standardowe.

Otrzymaliśmy bardzo wysoką wartość, co oznacza wyraźną korelację. Możemy przypuszczać, że jest tak dlatego, że dane dotyczą sportowców światowej klasy, którzy muszą posiadać odpowiednie parametry fizyczne, by móc startować w zawodach na tak wysokim poziomie. Stąd mamy niewiele odstających danych. W przypadku zwykłych ludzi rozrzut wartości byłby prawdopodobnie znacznie większy, przez co współczynnik korelacji byłby mniejszy.

5. Podsumowanie

Zgodnie z celem raportu zbadałem zbiór danych opisujący parametry fizyczne zawodników hokeja. Po przyjrzeniu się histogramom oraz wykresom gęstości empirycznej doszedłem do wniosku, że dane mogą pochodzić z rozkładu normalnego, co sprawdziłem porównując gęstość i dystrybuantę empiryczną z gęstością i dystrybuantą teoretyczną rozkładu normalnego. Jako parametry podstawiłem wyliczone wartości średnie oraz wariancje. Okazało się, że rzeczywiście rozkład naszych danych wpasowuje się w rozkład normalny.

Następnie sprawdziłem w jakim stopniu waga i wysokość ze sobą korelują. Otrzymana wartość współczynnika Pearsona wskazuje na znaczną korelację liniową między zbiorami. Stąd wniosek jest taki, że większa wysokość daje lepsze predyspozycje do osiągnięcia większej wagi, co jest w tym sporcie zapewne pożądane, by zawodnik mógł przewracać zawodników drużyny przeciwnej, a nie być przewracanym przez nich.

Źródła

[1] https://figshare.com/articles/dataset/Height_of_ice_hockey_players/3394735/2