RAPORT NR 1 | 15 maja 2022

Analiza danych statystycznych na przestrzeni dekad w koszykówce mężczyzn

Julia Grzegorzewska 262314, Wiktoria Fimińska 262283

STATYSTYKA STOSOWANA

dr Aleksandra Grzesiek

wtorek 7:30

Spis treści

1. Wstęp

1.1. Podstawy teoretyczne

Współcześnie **statystykę** można najprościej opisać jako zbiór danych dotyczących jakiegoś zjawiska lub procesu. To także wszystkie prace wykonane na tych danych, obliczanie charakterystyk, a przede wszystkim jest to nauka zajmująca się obserwacją zjawisk, wyciąganiem wniosków i eksperymentowaniem w celu potwierdzania swych założeń. Końcowym etapem badania statystycznego jest interpretacja zebranych danych i ich analiza. [1]

Podstawowymi pojęciami statystycznymi są: **zbiorowość statystyczna** - zbiór podobnych do siebie pod względem jakiejś cechy danych, **próba** - podzbiór danych, który poddany jest badaniu statystycznemu, **cecha statystyczna** - własność obiektu, który tworzy zbiorowość statystyczną, możemy je podzielić na mierzalne i niemierzalne, **rozmiar próby** - ilość obserwacji.[2]

1.2. Cel pracy

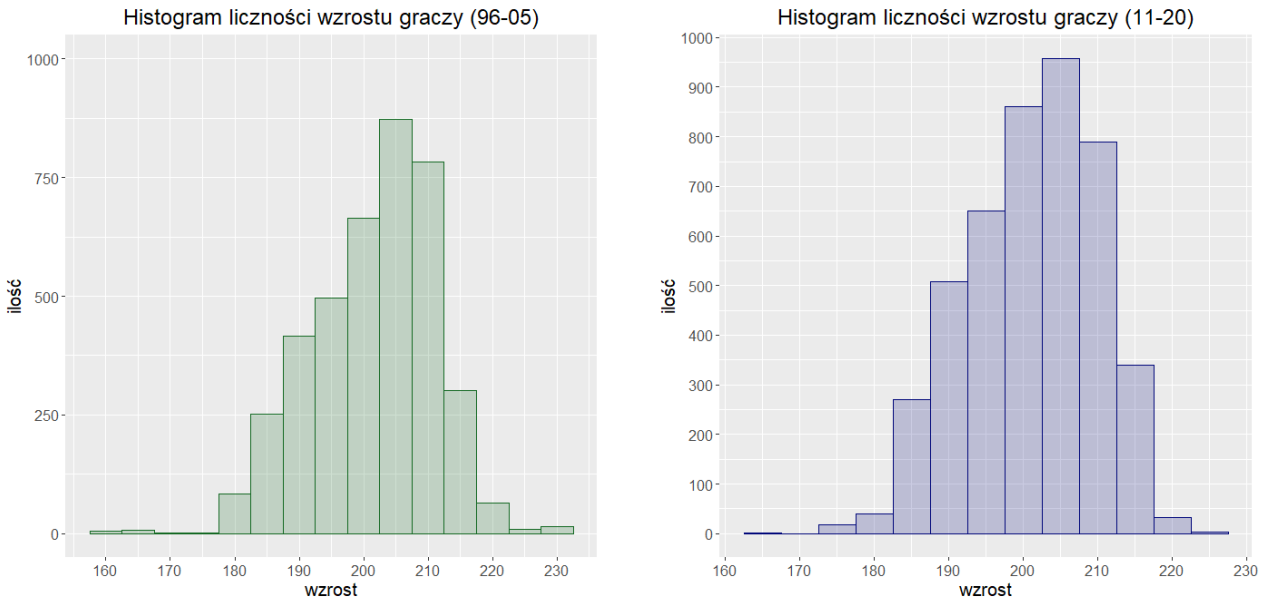
Celem niniejszej pracy jest przeprowadzenie analizy danych dotyczących wzrostu graczy ligi koszykarskiej NBA [3] na przestrzeni dwóch dekad (lata 1996-2005 oraz 2011-2020) i przedstawienie ich wizualizacji . Liczba zawodników wzrastała na przestrzeni lat, z tego powodu w pierwszej próbce jest 3972 danych, a w drugiej o 495 więcej. Nie wpłynęło to jednak znacząco na wiarygodność analizy. Wzrost zawodników wyrażony jest w centymetrach. Wszystkie obliczenia i analizy wykonywane były w języku R, z wykorzystaniem bibliotek *ggplot2* i *moments*.

2. Wizualizacja danych

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

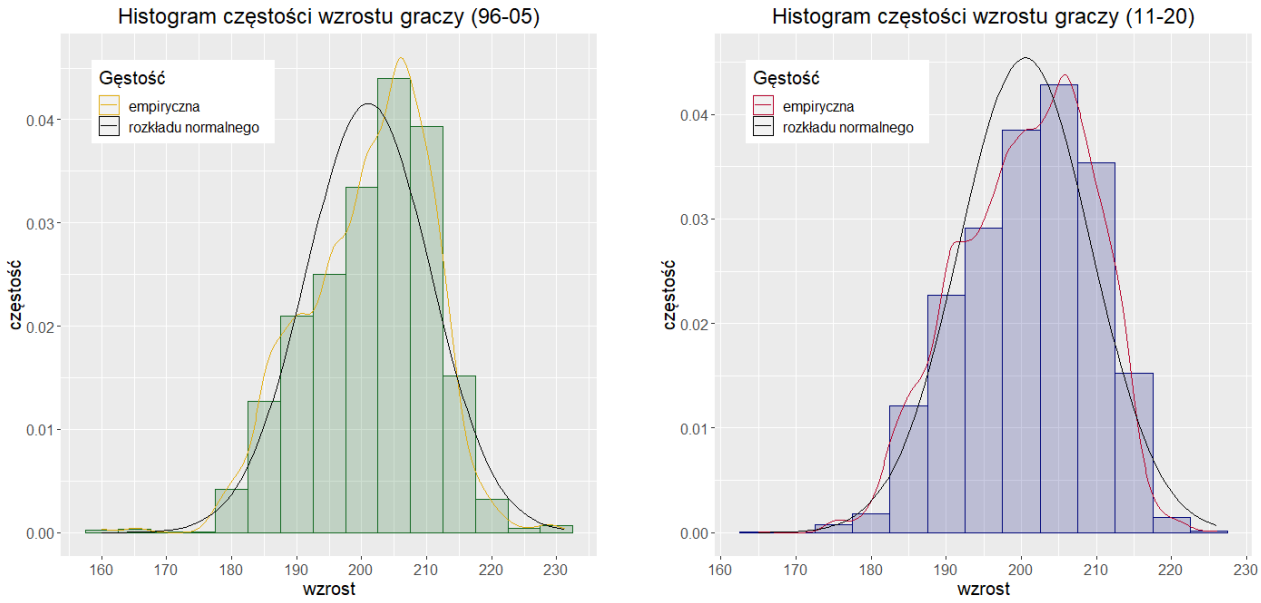
2.1. Histogramy

Pierwszym ze sposobów wizualizacji danych jest histogram liczności, który obrazuje liczebność występowania konkretnego wzrostu w badanej próbie.



Rysunek 1

Aby móc porównać oba zestawy danych wykonano histogramy częstości, które w istocie są empirycznymi odpowiednikami gęstości rozkładu. Naniesione zostały także wykresy gęstości rozkładu normalnego, aby przekonać się, czy pokrywają się one z gęstościami naszych prób.

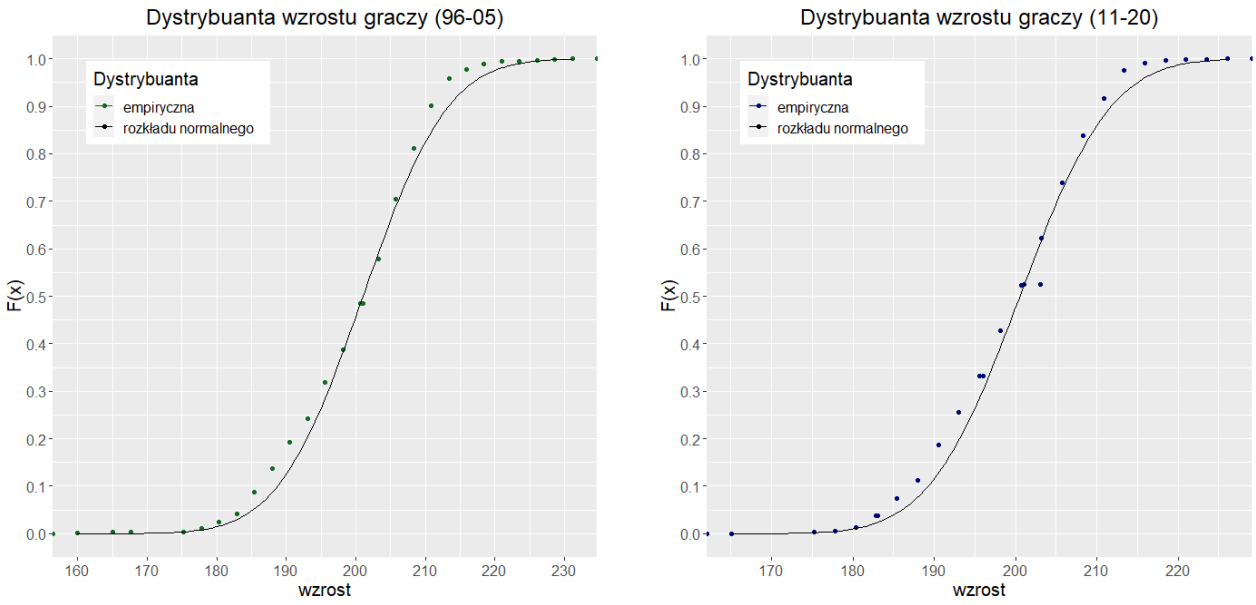


Rysunek 2

Z histogramów można odczytać modę, czyli przedział, do którego wpada najwięcej wartości. W obydwu dekadach moda wynosi $[202.5, 207.5)$, co świadczy o tym, że przeciętny wzrost zawodnika utrzymywał się na stałym poziomie. Można również zauważyć, że nie odnotowano drastycznych zmian, co do rozkładu wzrostu koszykarzy. Jedynie w przedziale $[192.5, 202.5)$ w latach 11-20 pojawił się niewielki wzrost częstości w porównaniu z okresem wcześniejszym. Patrząc na gęstości rozkładu normalnego oraz empiryczne w obu przypadkach wyraźnie widoczne są różnice między nimi, a co za tym idzie nasze dane nie pochodzą z rozkładu normalnego. Odczytać można również informację odnośnie asymetrii histogramów i rozkładów, która jest w obu dekadach ujemna - rozkład jest lewostronnie skośny.

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. Wykonano również wykres dystrybuanty rozkładu normalnego, aby ponownie sprawdzić ich zależność.

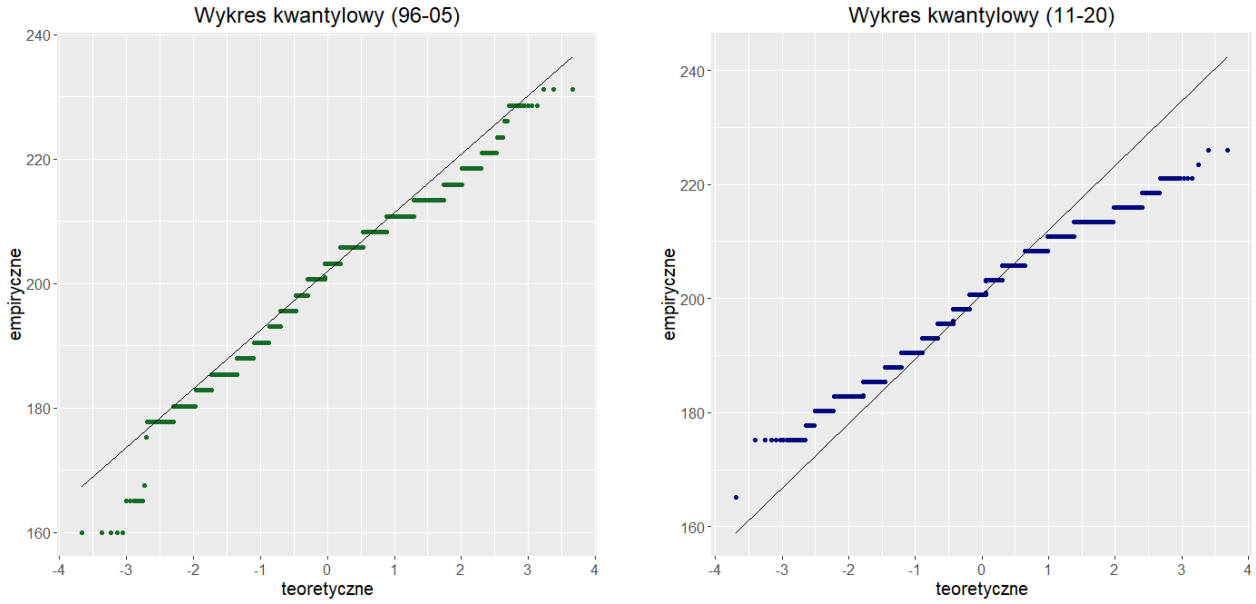


Rysunek 3

Patrząc na oba wykresy widać ich znaczące podobieństwo, co potwierdza tezę, że rozkład danych pozostawał niezmienny na przestrzeni rozpatrywanych dwóch dekad. Zauważyć można również, że dystrybuanty empiryczne odbiegają od teoretycznych, więc nie jest to próba z rozkładu normalnego.

2.3. Wykresy kwantylowe

Aby upewnić się, że dane na pewno nie pochodzą z rozkładu normalnego wykonane zostały wykresy kwantylowe, które są graficzną metodą testowania zgodności danych empirycznych z rozkładem teoretycznym.



Rysunek 4

W obu przypadkach wyraźnie widać rozbieżność kwantyli rozkładu normalnego z kwantylami empirycznymi badanej próby, co potwierdza, że nie pochodzi ona z tego rozkładu.

3. Statystyki opisowe

Ważnym elementem analizy danych rzeczywistych są metody i statystyki opisowe, które również w pewien sposób opisują właściowości badanej próby i pozwalają na ich podsumowanie oraz wyciągnięcie odpowiednich wniosków. Jedną z takich metod jest wyznaczanie miar rozkładu, które dostarczają informacji o charakterze badanego rozkładu. W dalszej części pracy n będzie oznaczało rozmiar próby.

3.1. Miary położenia

Miary położenia wskazują wokół jakich wartości oscylują analizowane dane. Wyróżnia się miary klasyczne, takie jak średnie oraz pozycyjne, czyli na przykład kwartyle. Obie te miary opisują pewne własności i cechy próby z innego punktu widzenia.

3.1.1. Średnie

Podstawowe rodzaje średnich to:

— średnia arytmetyczna

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;$$

— średnia harmoniczna

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}};$$

— średnia geometryczna

$$\bar{x}_g = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}};$$

— średnia ucinana

$$\bar{x}_u = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_i;$$

— średnia winsorowska

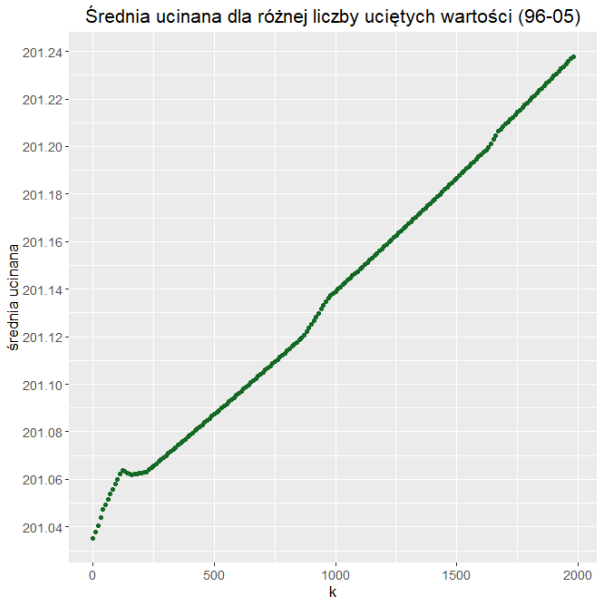
$$\bar{x}_w = \frac{1}{n}\left[(k+1)x_{k+1} + \sum_{i=k+2}^{n-k-1} x_i + (k+1)x_{n-k}\right].$$

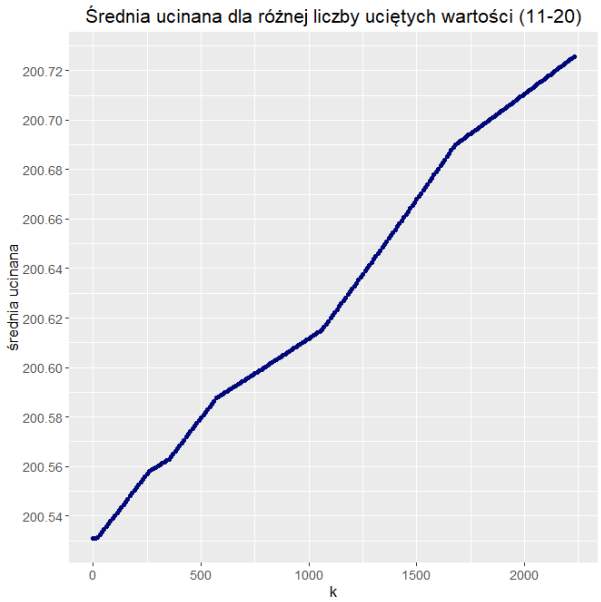
W tabeli 1 oraz na rysunkach 5 i 6 przedstawiono wartości powyższych miar dla badanych zestawów danych.

	96-05	**11-21**
$\bar{x}$	201.032	200.529
$\bar{x}_h$	200.563	200.137
$\bar{x}_g$	200.800	200.334

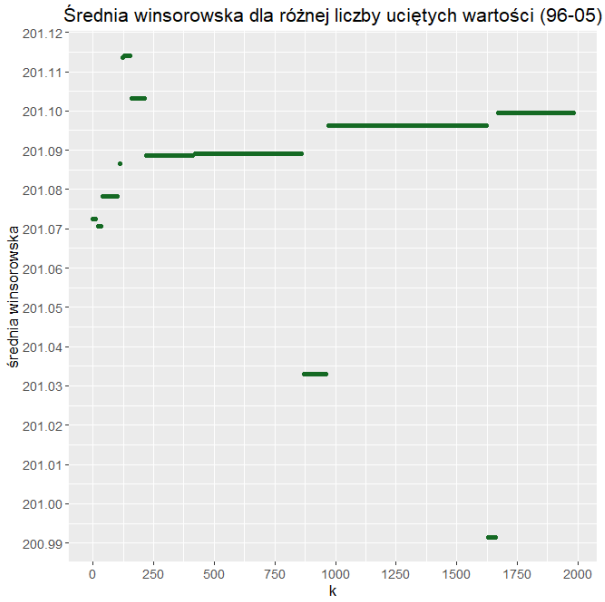
Tabela 1: Średnie

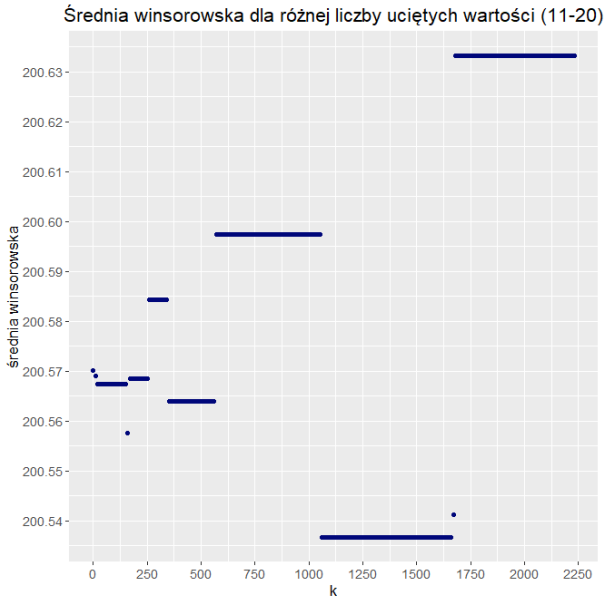
Patrząc na średnią arytmetyczna, harmoniczną i geometryczną widać, że w każdej z dekad są one do siebie zbliżone. Porównując oba zestawy danych ze sobą, również można zauważyć ich podobieństwo. W latach 1996-2005 średnie te są tylko nieznacznie wyższe niż w 2011-2021. Świadczy to ponownie o tym, że przeciętny wzrost zawodnika pozostawał na stałym poziomie.





Rysunek 5





Rysunek 6

Wartości w średniej ucinanej zmieniają się zaledwie o 0.2, gdy odrzucamy 2000 próbek obserwacji ekstremalnych. Oznacza to, że w danych nie ma zbyt wiele wartości odstających, które mogą zakłócać ich analizowanie. Dzieje się tak zarówno w latach 1996-2005, jak i 2011-2020. Średnie ucinane rosną w obu przypadkach, co świadczy lewoskośności prób. W średniej winsorowskiej różnice wartości w obu przypadkach są rzędu 0.1, co tylko potwierdza tezę postawioną wyżej przy średniej ucinanej. Porównując próbki możemy zauważyć, że w pierwszej dekadzie wartości te zachowują się trochę stabilniej i nie ma takiego rozstrzału na wykresach i zmian w wynikach. Oscylują one jednak wokół bardzo zbliżonych wartości, więc różnice te są znikome. Średnie winsorowskie przy coraz większej wartości współczynnika k dążą do median.

3.1.2. Kwartyle

Są to miary, które dzielą zbiór danych na cztery grupy. Wyróżniamy następujące kwartyle:

— drugi kwartyl $Q2$
Jest to szczególny kwartyl, nazywany inaczej medianą, który dzieli zbiór na dwa równoliczne podzbiory i wyraża się następującym wzorem

$$x_{med} = \begin{cases} x_{((n+1)/2)}, & \text{n nieparzyste} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{n parzyste.} \end{cases}$$

— pierwszy kwartyl $Q1$ (mediana grupy obserwacji mniejszych od Q2)
— trzeci kwartyl $Q3$ (mediana grupy obserwacji większych od Q2)

W tabeli 2 przedstawiono wartości tych miar dla badanych danych.

	96-05	**11-20**
x_{med}	203.20	200.66
$Q1$	195.58	193.04
$Q3$	208.28	208.28

Tabela 2: Kwartyle

Wyznaczone mediany są w obu dekadach większe od średnich arytmetycznych, co świadczy o lewoskośności tychże danych.

3.2. Miary rozproszenia

Miary rozproszenia opisują, jak zróżnicowane są wartości w danej próbie. Zaliczają się do nich:

— rozstęp międzykwartylowy
$$IQR = Q3 - Q1;$$

— rozstęp z próby
$$R = x_{(n)} - x_{(1)};$$

— wariancja
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 ;$$

— odchylenie standardowe
$$S = \sqrt{S^2};$$

— współczynnik zmienności
$$V = \frac{S}{\bar{x}} (\cdot 100\%)$$

Wartości powyższych miar przedstawiono w tabeli 3. Analizując wartości rozstępu międzykwartylowego można stwierdzić, że w latach 2011-2020 wzrost graczy był bardziej zróżnicowany. Rozstęp z próby, który pokazuje różnicę wzrostu między najniższym, a najwyższym zawodnikiem, w obu dekadach jest zaskakujący, bo wynosi aż 71.12 i 60.96 cm. Odchylenie standardowe mówiące o tym, o ile średnio odchylają się badane wartości od średniej arytmetycznej jest podobne dla badanych prób. Tak samo sytuacja wygląda dla współczynnika zmienności - w obu przypadkach wynosi około 4.5%.

	96-05	**11-20**
IQR	12.70	15.24
R	71.12	60.96
S^2	92.05	77.20
S	9.59	8.79
V	4.77%	4.38%

Tabela 3: Miary rozproszenia

3.3. Inne miary

— współczynnik skośności (miara asymetrii)

$$\alpha = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S} \right)^3 ;$$

— 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 4.

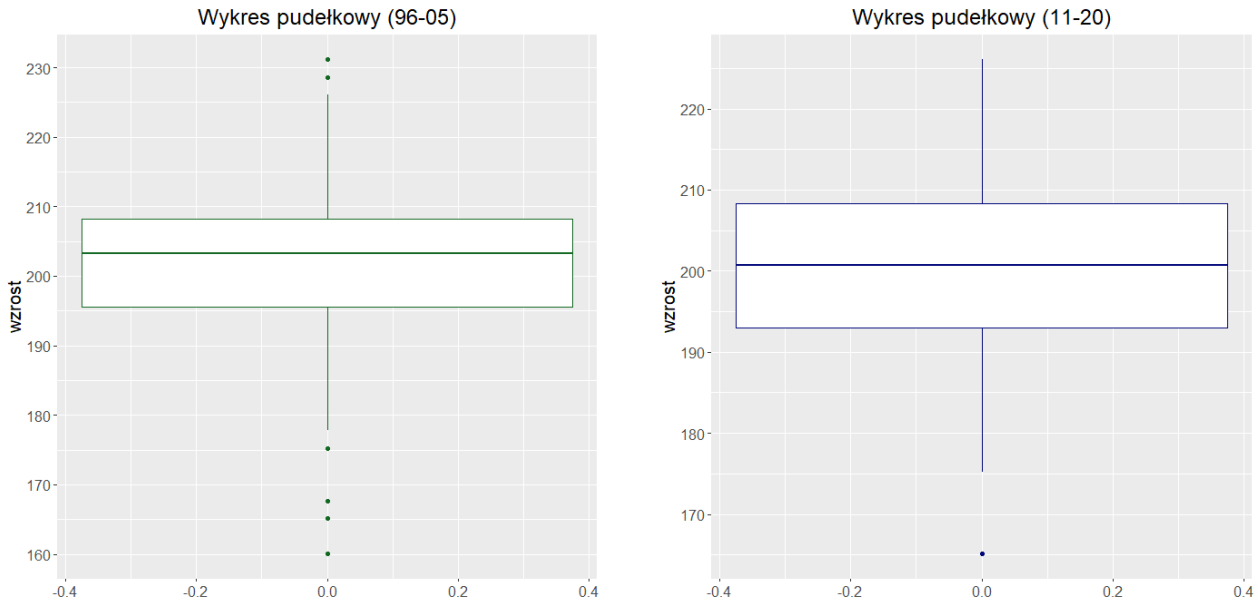
	96-05	**11-20**
α	-0.42	-0.31
K	3.16	2.51

Tabela 4: Miary asymetrii i spłaszczenia

Dla obu badanych prób współczynnik asymetrii jest ujemny, co świadczy o tym, że rozkłady te są lewostronnie skośne, co można było również zauważyć, analizując histogramy tychże zbiorów danych. Patrząc na kurtozę, która określa jak rozmieszczają się wartości wokół średniej, można stwierdzić, że rozkłady są skupione wokół swoich wartości oczekiwanych, co zostało dobrze zobrazowane na histogramach. Otrzymane wyniki ponownie potwierdzają, iż badane dane nie pochodzą z rozkładu normalnego, dla którego miara ta wynosi 3.

3.4. Wykresy pudełkowe

Aby móc zaprezentować graficznie omówione statystyki wykonano wykresy pudełkowe.



Rysunek 7

Z powyższych wykresów pudełkowych można odczytać wartości niektórych podstawowych statystyk. Pozioma linia, znajdująca się wewnątrz prostokąta oznacza medianę, krótsze boki prostokątów pokazują rozstęp międzykwartylowy, natomiast punkty obrazują wartości odstające, które mogą zaburzać interpretację wyników. Można zauważyć, że w pierwszej analizowanej dekadzie wartości odstających jest więcej niż w dekadzie drugiej, co pokazuje, że w tym okresie pojawiało się więcej zawodników o nietypowym dla tego sportu wzroście. W latach 1995-2005 mediana leży wyraźnie ponad średnią, co potwierdza, że rozkład tych danych jest lewostronnie skośny. W drugim zestawie danych nie jest to zbyt zauważalne. Mocno widoczna jest róznica wartości rozstępów międzykwartylowych - był on większy w latach 2011-2020. Ponownie jednak można stwierdzić, że badane rozkłady zachowują sie podobnie.

4. Podsumowanie

Badanym zestawem danych był wzrost graczy w lidze NBA na przestrzeni dwóch dekad. Dane te zostały najpierw poddane analizie graficznej; wykonano histogramy, wykresy gęstości, dystrybuant oraz kwantylowe. Następnie policzono znane statystyki opisowe, a wyniki zaprezentowano w tabelach. Wszystkie te czynności pozwoliły stwierdzić, że w ciągu ponad 20 lat wzrost graczy nie zmienił się znacząco. Różnica ilości próbek może pokazywać wzrost zainteresowania tym sportem i chęci rozwijania się w tejże dziedzinie. Możliwe, że gdyby dane pochodziły z większego przedziału czasowego, różnice te byłyby bardziej zauważalne.

Literatura

[1] https://pl.wikipedia.org/wiki/Statystyka#Statystyka_stosowana
[2] https://eks.stat.gov.pl/materialy/scenariusze/miary_statystyczne/materialy_dla_nauczyciela.pdf
[3] https://www.kaggle.com/code/justinas/nba-height-and-weight-analysis/data