**Sprawozdanie I**

Analiza danych statystycznych

Katarzyna Rybarczyk

1. **Informacje o danych**

Dane użyte do sprawozdania prezentują wiek, wzrost i wagę próbki 250 dorosłych mężczyzn ze Stanów Zjednoczonych. Zostały pobrane z amerykańskiej strony Data and Story Library (<https://dasl.datadescription.com/datafile/bodyfat/>). Jakość danych jest wysoka – nie istnieją puste obserwacje, o dokładności pomiarów zapewnia także autor. Wszystkie zmienne są liczbowe.

Oryginalne zestawienie zawierało więcej zmiennych, np. obwód uda, klatki piersiowej i innych części ciała, jednak nie zostały one użyte w raporcie – wybrano dane najłatwiejsze do interpretacji.

Waga została podana w funtach, więc pierwszym krokiem było przeliczenie jej na kilogramy (przyjęte zostało 1lbs = 0,45kg). Analogicznie został przeliczony wzrost z cali na centymetry   
(1’’ = 2,54cm).

Grupowanie zmiennych będzie odbywać się z uwagi na wiek – zbadane zostanie zatem zróżnicowanie rozkładów wzrostu i wagi dla grupy młodych dorosłych, osób w średnim wieku   
i osób starszych.

1. **Analiza opisowa**
   1. **Statystyki opisowe[[1]](#footnote-1)**

Jako pierwszy etap analizy sprawdzone zostały statystyki – miary położenia, rozproszenia, asymetrii i spłaszczenia.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Liczba obserwacji** | **Średnia** | **Wariancja** | **Odchylenie standardowe** | **Współczynnik zmienności** | **Min** | **Max** | **Q0.25** | **Q0.5** | **Q0.75** | **Skośność** | **Kurtoza** |
| **Wiek** | 250 | 44,88 | 160,08 | 12,65 | 28,19% | 22,00 | 81,00 | 35,25 | 43,00 | 54,00 | 0,28 | 2,55 |
| **Wzrost** | 250 | 178,57 | 44,17 | 6,65 | 3,72% | 162,56 | 197,49 | 173,36 | 177,80 | 183,52 | 0,11 | 2,57 |
| **Waga** | 250 | 80,14 | 148,01 | 12,17 | 15,18% | 53,33 | 118,24 | 71,33 | 79,26 | 88,54 | 0,37 | 2,80 |

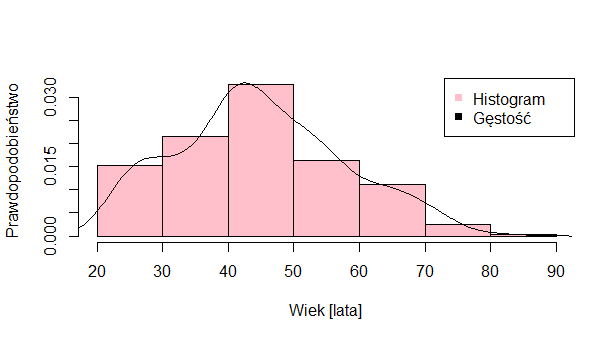
Obserwacje na podstawie statystyk:

* Próba zawiera różnorodny przekrój społeczeństwa – świadczy o tym szeroki zakres danych od minimum do maksimum.
* Rozkłady wszystkich zmiennych są delikatnie prawostronnie skośne – najbardziej symetryczny jest rozkład wzrostu (skośność najbliższa 0).
* Rozkłady wszystkich zmiennych są także delikatnie platokurtyczne (kurtoza < 3).
* Kształt rozkładu wagi najbliższy jest rozkładowi normalnemu (kurtoza najbliżej 3).
* Pomiary wzrostu cechują się najmniejszym rozproszeniem (mały współczynnik zmienności).
  1. **Rozkłady zmiennych**

Kolejnym etapem jest zobrazowanie rozkładów zmiennych za pomocą histogramu i estymatora gęstości.

Obraz zawierający tekst, mapa

Opis wygenerowany automatycznieObraz zawierający tekst, mapa

Opis wygenerowany automatycznieRozkład wieku nie przypomina innego znanego nam rozkładu. Rozkład z Rysunku 2 przypomina krzywą gaussowską, jednak biorąc pod uwagę kurtozę, wybrano rozkład wagi do porównania   
z normalnym o średniej i wariancji z rozkładu tej zmiennej. Można zauważyć podobieństwo.

Rysunek 3: Histogram wagi

Rysunek 2: Histogram wzrostu

Rysunek 1: Histogram wieku

Obraz zawierający zrzut ekranu

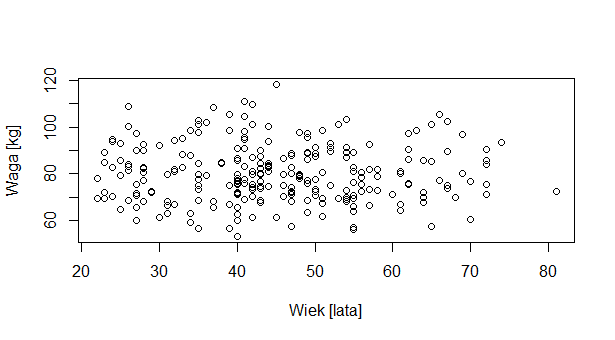
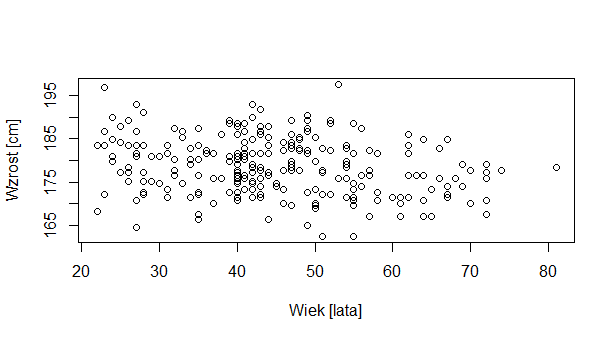
Opis wygenerowany automatycznieDo wizualizacji wspomnianych wcześniej statystyk stworzone zostały także wykresy pudełkowe.

Rysunek 4: Wykresy pudełkowe

Potwierdzają one poczynione wcześniej obserwacje, szczególnie niewielką asymetryczność

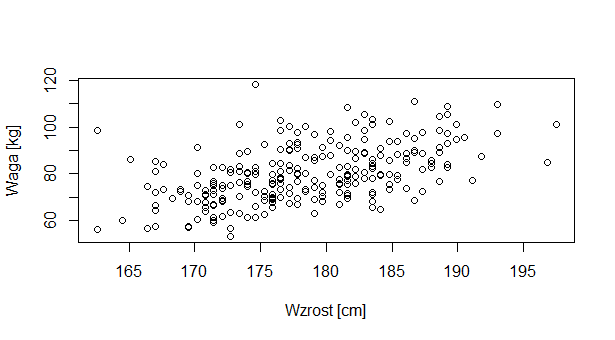
w lewą stronę (przesunięcie mediany względem średniej). Zauważyć można też brak większych wartości odstających.

* 1. **Zależności między zmiennymi**

Następnie na wykresach 2D umieszczono dwie z trzech zmiennych w celu wychwycenia ewentualnych zależności.

Rysunek 5: Zależność wagi od wieku

Rysunek 6: Zależność wzrostu od wieku

Na powyższych wykresach nie udało odnaleźć się zależności, jednak ciekawszy wynik przyniosło zestawienie zmiennych wagi i wzrostu:

Rysunek 7: Zależność wzrostu od wagi

Możemy dostrzec tutaj lekką zależność liniową. Policzone współczynniki korelacji Pearsona[[2]](#footnote-2) prezentują się następująco:

|  |  |
| --- | --- |
|  | **Współczynnik korelacji** |
| **Rysunek 5** | -0,0161 |
| **Rysunek 6** | -0,2459 |
| **Rysunek 7** | 0,5129 |

Poziom współczynnika korelacji powyżej 0,5 można uznać już za znaczący, a zmienne wzrostu i wagi traktować jako prawdopodobnie zależne (w sposób rosnący, ponieważ współczynnik > 0). Zależność ta nie jest bezpodstawna – znajduje faktyczne odzwierciedlenie w biologii. Biologicznie uwarunkowaną zależnością jest też spadek wzrostu wraz z wiekiem (Rysunek 6), jednak tutaj współczynnik nie jest tak znaczący.

1. **Analiza z podziałem na grupy**

Grupowanie nastąpiło z uwagi na wcześniej wspomniane trzy grupy wiekowe. W celu zachowania podobnych wielkości próbek, policzone zostały kwantyle rzędu 0.33 i 0.67, które wyniosły odpowiednio 40 i ok. 50. Taki podział wydał się jednak mało intuicyjny, a środkowy przedział za wąski, więc został rozszerzony kosztem większego zróżnicowania ilości obserwacji   
w każdym przedziale. Ostatecznie otrzymaliśmy zatem 63 obserwacje w przedziale poniżej 36 roku życia (młodzi dorośli), 111 obserwacji w przedziale 36-50 (osoby w wieku średnim) oraz 76 obserwacji w przedziale powyżej 50 roku życia (osoby starsze).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| WAGA | **Liczba obserwacji** | **Średnia** | **Wariancja** | **Odchylenie standardowe** | **Min** | **Maks** | **Q0.25** | **Q0.5** | **Q0.75** | **Skośność** | **Kurtoza** |
| **<36** | 63 | 80,14 | 152,42 | 12,35 | 56,59 | 108,79 | 69,81 | 80,55 | 89,66 | 0,16 | 2,21 |
| **36-50** | 111 | 80,76 | 150,53 | 12,27 | 53,33 | 118,24 | 73,18 | 79,31 | 87,69 | 0,53 | 3,36 |
| **>50** | 76 | 79,22 | 143,19 | 11,97 | 56,25 | 105,41 | 69,95 | 76,95 | 89,16 | 0,30 | 2,36 |

* 1. **Statystyki opisowe**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| WZROST | **Liczba obserwacji** | **Średnia** | **Wariancja** | **Odchylenie standardowe** | **Min** | **Maks** | **Q0.25** | **Q0.5** | **Q0.75** | **Skośność** | **Kurtoza** |
| **<36** | 63 | 179,71 | 46,62 | 6,83 | 164,47 | 196,85 | 174,94 | 180,34 | 183,83 | 0,03 | 2,62 |
| **36-50** | 111 | 179,54 | 38,86 | 6,23 | 165,10 | 193,04 | 174,94 | 179,71 | 184,47 | 0,04 | 2,22 |
| **>50** | 76 | 176,21 | 43,02 | 6,56 | 162,56 | 197,49 | 171,45 | 175,90 | 180,18 | 0,45 | 3,39 |

Obserwacje na podstawie statystyk:

* Położenie rozkładów wagi jest dość podobne dla każdej z grup.
* Wśród rozkładów wagi największą skośność (prawostronną) oraz kurtozę (leptokurtyczność) wykazuje rozkład grupy 36-50.
* Położenie grupy >50 wśród rozkładów wzrostu jest widocznie przesunięte w lewo (mniejsza średnia i kwantyle). Powodem może być wspomniana wcześniej zależność wzrostu od wieku. Ta grupa ma również największą skośność (prawostronną) i kurtozę (> 3) ze wszystkich rozkładów wzrostu.
* Podzielenie wzrostu na grupy spowodowało oddalenie się od rozkładu normalnego.
  1. **Rozkłady zmiennych**

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznieObraz zawierający zrzut ekranu

Opis wygenerowany automatycznieWykresy pudełkowe reprezentujące powyższe informacje:

Rysunek 8: Wykresy pudełkowe

Zauważalne jest wspomniane wyżej przesunięcie w dół rozkładu wzrostu dla grupy >50   
w stosunku do innych.

Histogramy poszczególnych zmiennych w odpowiednich grupach:



Rysunek 9: Histogramy pogrupowanych zmiennych

Rozkłady wzrostu utrzymują gaussowskie kształty, lecz oddaliły się od rozkładu normalnego. Rozkłady wagi mają podobne kształty w grupie 36-50 i >50, a w grupie <36 są bardziej rozproszone.

* 1. **Zależność zmiennych**

Podział na grupy nie wpłynął na zależność zmiennych, dlatego nie zostały zamieszczone tutaj kolejne wykresy. Współczynniki korelacji prezentują się następująco:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Wiek/Waga** | **Wiek/Wzrost** | **Wzrost/Waga** |
| **<36** | -0,0074 | -0,2732 | 0,5325 |
| **36-50** | -0,0069 | -0,0390 | 0,5209 |
| **>50** | 0,0956 | -0,1646 | 0,4978 |

Co ciekawe, domniemana zależność wieku i wzrostu jest najsilniejsza w grupie poniżej 36 roku życia i bardzo słaba dla osób w grupie średniej.

1. **Wzory i definicje[[3]](#footnote-3)**
2. Średnia arytmetyczna
3. Wariancja
4. Współczynnik zmienności
5. Kwantyl rzędu r

oraz

1. Skośność
2. Kurtoza

( – czwarty moment centralny)

1. Współczynnik korelacji Pearsona

Współczynnik ten przyjmuje wartości z zakresu [-1, 1], gdzie -1 to dokładna ujemna zależność liniowa, 0 to brak korelacji, a 1 to dodatnia zależność liniowa. Duże wartości odstające mogą mocno wpłynąć na ten współczynnik.

1. **Wnioski**

Dane użyte do raportu są dobre jakościowo – nie ma pustych obserwacji, elementy próby są różnorodne, brak dużych wartości odstających. Rozkłady wagi i wzrostu w społeczeństwie przypominają rozkład normalny, lecz są delikatnie lewostronnie skośne i platokurtyczne. Udało się zaobserwować możliwą liniową zależność dodatnią między zmienną wzrostu i wagi, co znajduje swoje wytłumaczenie w biologii. Możliwa jest także zależność ujemna między wiekiem a wzrostem, jednak nie można jednoznacznie jej określić.

Podział danych na grupy według wieku ujawnił zależność wzrostu od wieku – położenie zmiennej wzrostu w grupie >50 roku życia było znacząco obniżone względem innych grup. Można wysunąć hipotezę o podłożu biologicznym, iż zależność ta pojawia się dopiero od pewnego momentu życia.

1. Wzory na poszczególne statystyki znajdują się na końcu sprawozdania [↑](#footnote-ref-1)
2. Definicja i wzór znajdują się na końcu sprawozdania [↑](#footnote-ref-2)
3. Punkty a) – f) z notatek z wykładów prof. Burneckiego,

   g) – https://pl.wikipedia.org/wiki/Wsp%C3%B3%C5%82czynnik\_korelacji\_Pearsona [↑](#footnote-ref-3)