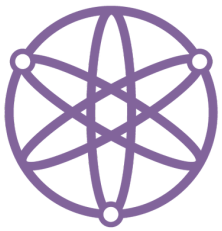


WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

Natalia Szczupak

Analiza wagi mieszańców Meksyku, Peru i Kolumbii

Projekt z przedmiotu Statystyczna Analiza Danych

Opiekun pracy:
dr Mariusz Startek, adiunkt w grupie pracowników dydaktycznych

Rzeszów, 2025

Spis treści

Opis użytych funkcji środowiska R

W trakcie realizacji projektu wykorzystano szereg funkcji dostępnych w języku R, zarówno z pakietu bazowego, jak i z zewnętrznego pakietu `e1071`. Poniżej przedstawiono zestawienie użytych funkcji wraz z krótkim opisem ich działania oraz wskazaniem pakietu, z którego pochodzą.

- `read.csv()` – wczytuje dane z pliku CSV do ramki danych (ang. data frame).
  Pakiet: base

- `mean()` – oblicza średnią arytmetyczną dla wektora liczbowego.
  Pakiet: base

- `sd()` – oblicza odchylenie standardowe dla wektora liczbowego.
  Pakiet: base

- `quantile()` – oblicza kwantyle (np. kwartyle, mediana, min, max) dla danego zbioru danych.
  Pakiet: base

- `var()` – oblicza wariancję wektora liczbowego.
  Pakiet: base

- `hist()` – tworzy histogram, czyli wykres słupkowy przedstawiający rozkład danych liczbowych.
  Pakiet: graphics

- `boxplot()` – tworzy wykres pudełkowy, prezentujący rozkład danych z uwzględnieniem mediany, kwartylów oraz wartości odstających.
  Pakiet: graphics

- `plot()` – funkcja ogólna do rysowania różnych typów wykresów, w tym dystrybuanty empirycznej oraz wykresu gęstości.
  Pakiet: graphics

- `ecdf()` – tworzy funkcję rozkładu empirycznego (dystrybuantę). Zwraca funkcję, którą można przekazać do `plot()`.
  Pakiet: stats

- `density()` – estymuje funkcję gęstości rozkładu zmiennej ciągłej metodą jądrową. Wynik przekazywany jest do funkcji `plot()`.
  Pakiet: stats
- `t.test()` – wykonuje test t-Studenta (jednopróbkowy, dwupunktowy lub dla dwóch prób zależnych/niezależnych). W projekcie wykorzystano go do testu średniej.
  Pakiet: stats
- `prop.test()` – wykonuje test proporcji dla jednej lub dwóch prób. W projekcie użyto go do sprawdzenia, czy udział osób otyłych różni się od założonego 30%.
  Pakiet: stats
- `skewness()` – oblicza współczynnik asymetrii (skośności) rozkładu.
  Pakiet: e1071
- `library()` – ładuje dodatkowy pakiet. W projekcie użyto tej funkcji do załadowania pakietu `e1071`, który zawiera funkcję `skewness()`.
  Pakiet: base

1. Wstęp

Celem niniejszego projektu jest analiza danych dotyczących otyłości z wykorzystaniem środowiska R. Analiza została przeprowadzona na zbiorze danych Obesity Prediction Dataset, udostępnionym na platformie Kaggle

(https://www.kaggle.com/datasets/adeniranstephen/obesity-prediction-dataset). Dane zawierają informacje o cechach stylu życia, nawykach żywieniowych oraz podstawowych parametrach fizycznych osób, takich jak waga czy wzrost, wraz z przypisaną diagnozą dotyczącą poziomu otyłości.

W ramach projektu przeprowadzono:

- opis wykorzystanych danych, wraz z uzasadnieniem ich wyboru;
- wyznaczenie podstawowych statystyk opisowych, takich jak średnia, odchylenie standardowe, kwartyle, rozstęp, współczynnik zmienności, współczynnik asymetrii oraz odchylenie ćwiartkowe;
- graficzną prezentację danych w postaci histogramu, wykresu pudełkowego, dystrybuanty empirycznej oraz wykresu gęstości;
- weryfikację dwóch hipotez statystycznych – test średniej oraz test proporcji;
- komentarz do uzyskanych wyników oraz listing użytego kodu;
- opis wykorzystanych funkcji środowiska R, z uwzględnieniem pochodzenia z odpowiednich pakietów.

Projekt ma na celu nie tylko zaprezentowanie umiejętności analizy danych w środowisku R, ale także praktyczne zastosowanie wiedzy statystycznej do rzeczywistego problemu z zakresu zdrowia publicznego.

2. Opis użytych danych

W projekcie wykorzystano zbiór danych pt. Obesity Prediction Dataset. Dane zawarte w tym zbiorze dotyczą czynników wpływających na poziom otyłości u ludzi, takich jak: nawyki żywieniowe, aktywność fizyczna, spożycie kalorii, a także podstawowe cechy antropometryczne.

Zbiór danych zawiera zarówno dane surowe, jak i syntetyczne, wygenerowane na podstawie wcześniej zebranych informacji. Dane obejmują m.in.:

- wiek, wzrost i wagę respondentów,
- częstotliwość aktywności fizycznej,
- nawyki żywieniowe, w tym np. spożywanie fast foodów, ilość posiłków dziennie, spożycie warzyw i picie wody,
- informacje o stylu życia, takie jak używki czy godziny spędzane przed ekranem,
- zmienną docelową (NObeyesdad), klasyfikującą osoby według poziomu otyłości (np. Normal_Weight, Overweight_Level_I, Obesity_Type_II itd.).

Wybór tego zbioru danych został podyktowany jego bogactwem informacyjnym, aktualnością problemu otyłości we współczesnym społeczeństwie oraz możliwością zastosowania różnorodnych metod statystycznych i wizualizacji w analizie danych. Analiza skupia się głównie na zmiennej waga (Weight), której rozkład oraz charakterystyki statystyczne są przedmiotem dalszych etapów projektu.

3. Analiza statystyczna danych

3.1. Parametry opisowe danych

W celu scharakteryzowania zmiennej „waga” (Weight) obliczono podstawowe parametry statystyki opisowej. Poniżej przedstawiono wybrane miary, które pozwalają ocenić rozkład wartości w zbiorze danych:

- **Średnia arytmetyczna**

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

  Średnia wynosi: **86,59 kg**. Oznacza to przeciętną wagę w badanym zbiorze.

- **Odchylenie standardowe**

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

  Odchylenie standardowe wynosi: **26,19**. Pokazuje przeciętne odchylenie wartości od średniej.

- **Wariancja**

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

  Wariancja wynosi: **685,98**. Jest to miara zmienności będąca kwadratem odchylenia standardowego.

- **Rozstęp (zakres)**

$$R = x_{\max} - x_{\min}$$

  Rozstęp wynosi: **134 kg** (173 kg – 39 kg). Wskazuje na rozpiętość danych.

- **Odchylenie ćwiartkowe (QD)**

$$QD = \frac{Q_3 - Q_1}{2}$$

  Wartość odchylenia ćwiartkowego: **20,98**. Miara rozrzutu środkowych 50% danych.

- **Kwantyle**
  - Min: 39,00 kg
  - Q1 (25%): 65,47 kg
  - Mediana (Q2): 83,00 kg

- Q3 (75%): 107,43 kg
- Max: 173,00 kg

Kwantyle nie wymagają wzoru, gdyż opierają się na pozycyjnych miarach uporządkowanego szeregu danych.

- **Współczynnik zmienności**

$$V = \frac{s}{\bar{x}}$$

Wartość współczynnika: **0,302**. Pokazuje względną zmienność danych względem średniej.

- **Współczynnik asymetrii (skośność)**

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

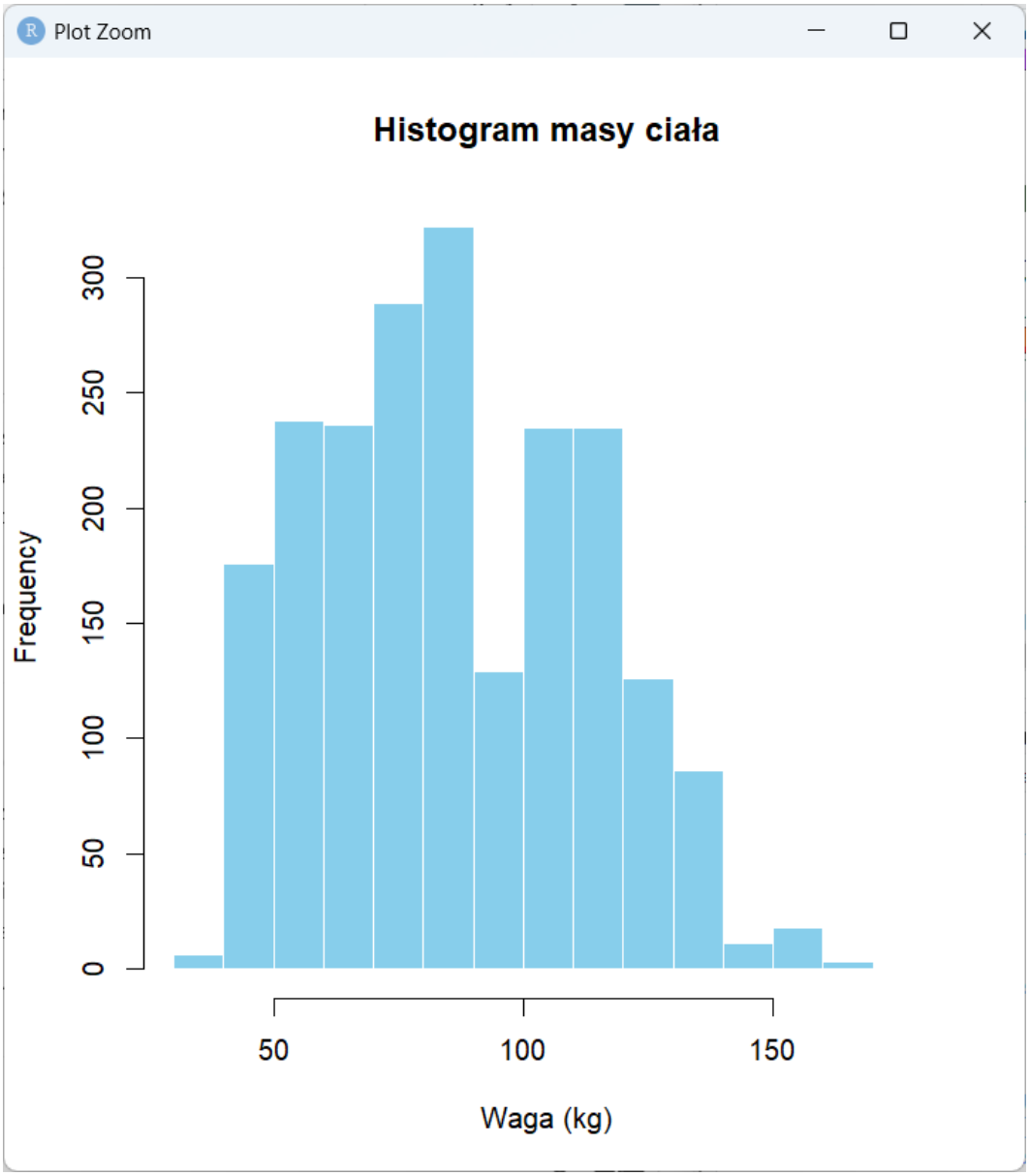
Wartość współczynnika asymetrii: **0,26**. Oznacza lekką prawostronną asymetrię rozkładu.

3.2. Graficzna prezentacja danych

W celu lepszego zobrazowania rozkładu zmiennej „waga" zastosowano cztery różne formy graficznej prezentacji danych. Wizualizacje te pozwalają na intuicyjne uchwycenie struktury rozkładu, występowania wartości odstających, kształtu gęstości oraz funkcji rozkładu empirycznego.

3.2.1. Histogram rozkładu wagi

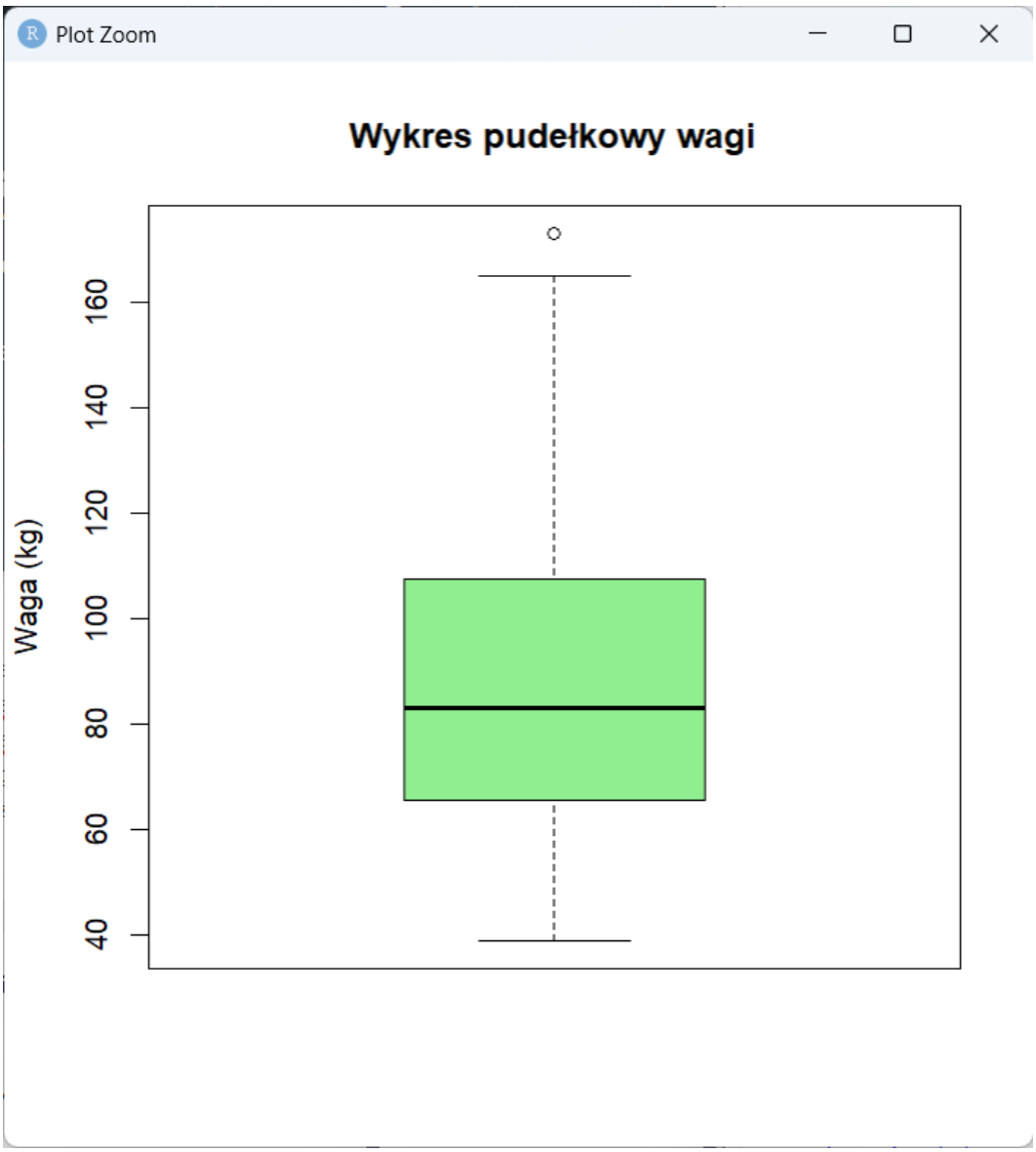
Na wykresie 3.1 przedstawiono histogram zmiennej „waga". Histogram pokazuje liczbę obserwacji w poszczególnych przedziałach klasowych. Można zauważyć, że rozkład nie jest idealnie symetryczny – występuje lekkie przesunięcie w prawo (skośność dodatnia), co pokrywa się z wcześniej obliczonym współczynnikiem asymetrii.



Rys. 3.1. Histogram rozkładu masy ciała

3.2.2. Wykres pudełkowy (boxplot)

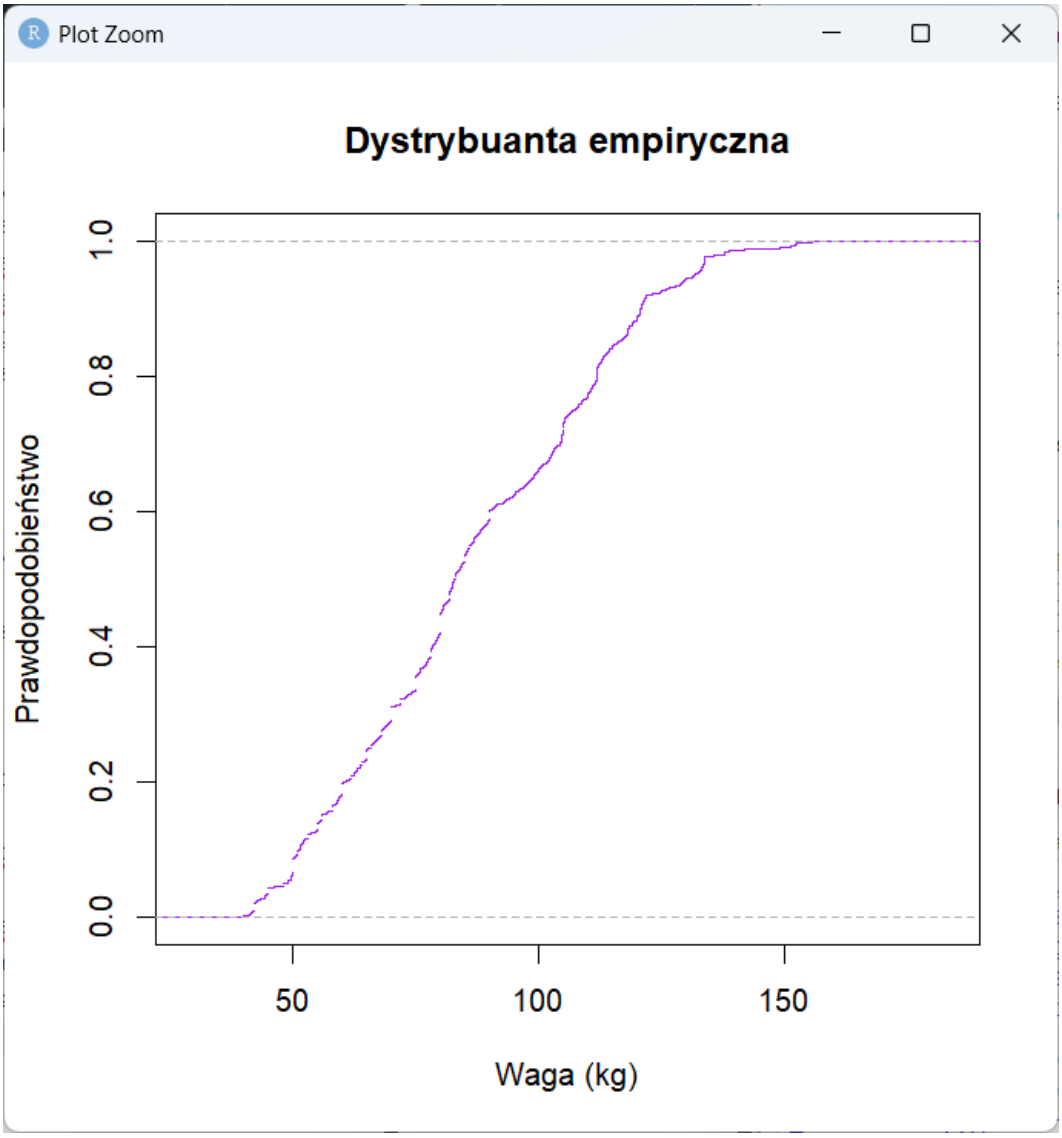
Na wykresie 3.2 zaprezentowano wykres pudełkowy (boxplot), który przedstawia medianę, kwartyle, rozstęp oraz potencjalne wartości odstające. Widać, że dane są nieco rozciągnięte w kierunku wyższych wartości, co potwierdza obserwowaną asymetrię i szeroki rozstęp (134 kg).



Rys. 3.2. Wykres pudełkowy zmiennej waga

3.2.3. Dystrybuanta empiryczna (ECDF)

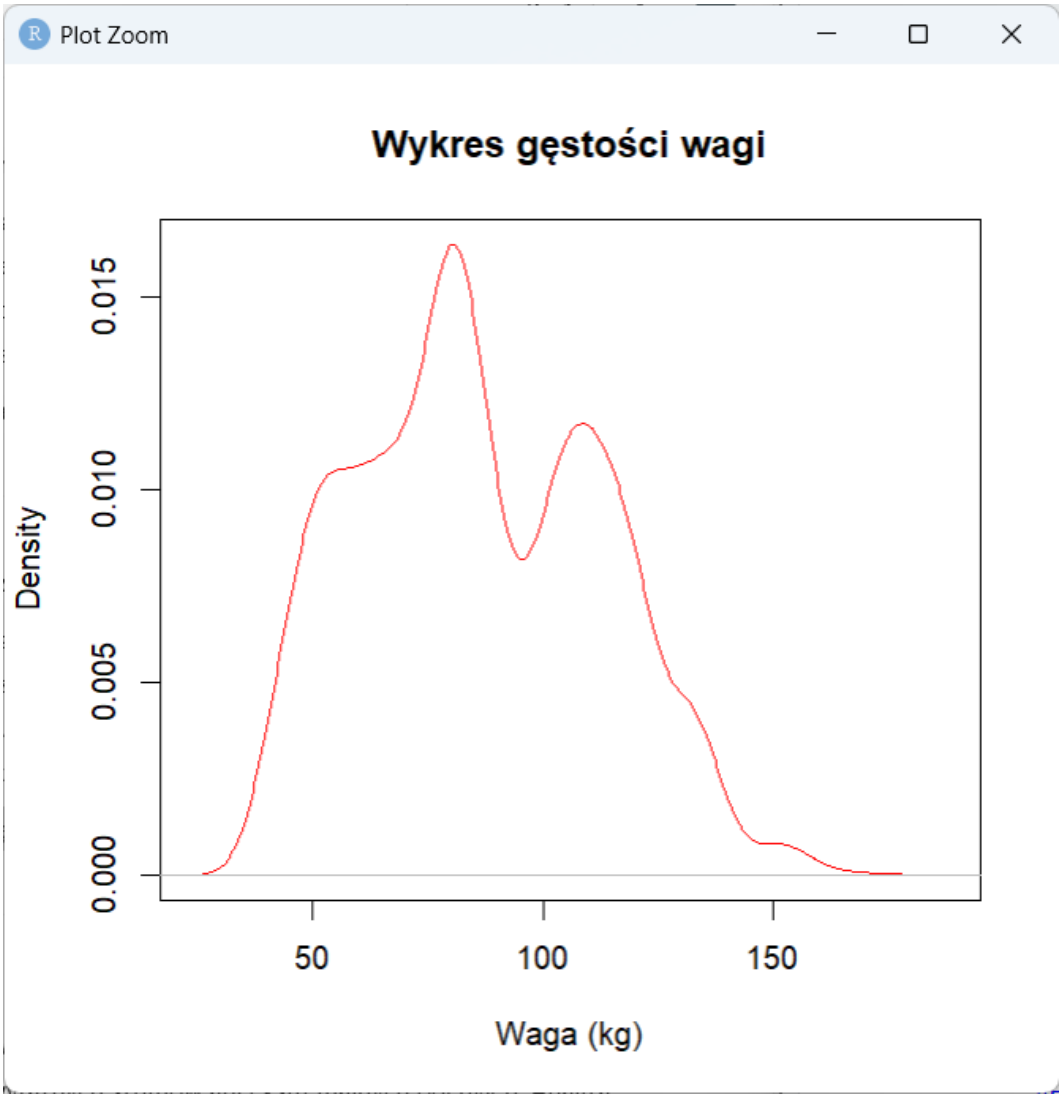
Wykres 3.3 przedstawia funkcję rozkładu empirycznego (dystrybuantę), która pokazuje skumulowane prawdopodobieństwo wystąpienia wag mniejszych lub równych danej wartości. ECDF pozwala zauważyć, że większość danych mieści się w przedziale 60–110 kg, a wzrost funkcji jest łagodny.



Rys. 3.3. Dystrybuanta empiryczna dla zmiennej waga

3.2.4. Wykres gęstości rozkładu

Na wykresie 3.4 przedstawiono wykres gęstości, który służy do estymacji rozkładu zmiennej ciągłej. W przeciwieństwie do histogramu, nie zależy on od podziału na klasy. Gęstość rozkładu jest zbliżona do rozkładu normalnego, lecz z wyraźnym ogonem po stronie wyższych wartości.



Rys. 3.4. Wykres gęstości dla zmiennej waga

Wszystkie zaprezentowane wykresy wskazują, że zmienna „waga" ma rozkład lekko skośny w prawo, bez wyraźnych szumów ani ekstremalnych odchyleń. Analiza graficzna potwierdza wnioski wyciągnięte wcześniej na podstawie statystyk opisowych.

3.3. Weryfikacja hipotez statystycznych

W celu sprawdzenia założeń dotyczących badanej populacji przeprowadzono dwa testy statystyczne: test średniej oraz test proporcji. Analizy wykonano przy standardowym poziomie istotności $\alpha = 0,05$.

3.3.1. Test średniej – czy średnia waga różni się od 70 kg?

Hipotezy:

$$H_0 : \mu = 70 \quad \text{(średnia waga wynosi 70 kg)}$$

$$H_1 : \mu \neq 70 \quad \text{(średnia waga różni się od 70 kg)}$$

Wyniki testu:

- wartość statystyki testowej: $t = 29{,}096$
- liczba stopni swobody: $df = 2110$
- p-wartość: $p < 2{,}2 \times 10^{-16}$
- przedział ufności dla średniej: [85,47; 87,70]
- średnia w próbie: $\bar{x} = 86{,}59$

Wniosek: Ponieważ p-wartość jest znacznie mniejsza niż poziom istotności $\alpha = 0{,}05$, **odrzucamy hipotezę zerową**. Oznacza to, że średnia masa ciała badanej populacji istotnie różni się od 70 kg — i wynosi przeciętnie około 86,6 kg.

2. Test proporcji – czy 30% osób to osoby otyłe?

Celem testu proporcji jest sprawdzenie, czy odsetek osób zakwalifikowanych jako otyłe (*Obesity_Type_I, II, III*) różni się od 30%.

Hipotezy:

$$H_0 : p = 0{,}3 \quad \text{(30\% populacji to osoby otyłe)}$$

$$H_1 : p \neq 0{,}3 \quad \text{(odsetek otyłych różny od 30\%)}$$

Wyniki testu:

- liczba otyłych: $x = 973$
- liczba wszystkich obserwacji: $n = 2111$
- oszacowana proporcja: $\hat{p} = 0{,}460$
- p-wartość: $p < 2{,}2 \times 10^{-16}$
- 95% przedział ufności: [0,439; 0,482]
- statystyka testowa: $\chi^2 = 258{,}01$

Wniosek: Ze względu na bardzo małą p-wartość odrzucamy hipotezę zerową. Oznacza to, że udział osób otyłych w badanej próbie istotnie różni się od 30%. W rzeczywistości osoby otyłe stanowią około **46%** badanej populacji.

Podsumowanie: Obie analizy doprowadziły do odrzucenia hipotez zerowych. Zarówno średnia waga, jak i udział osób otyłych w badanej próbie różnią się istotnie od zakładanych wartości (odpowiednio 70 kg oraz 30%).

4. Podsumowanie i wnioski końcowe

Celem niniejszego projektu była analiza danych dotyczących masy ciała osób pochodzących z obszarów Meksyku, Peru i Kolumbii, z wykorzystaniem środowiska R. Analizowano zmienną *waga* (Weight), będącą jednym z głównych parametrów w zbiorze danych „Obesity Prediction Dataset", dostępnym na platformie Kaggle.

W pierwszym etapie projektu przeprowadzono szczegółową charakterystykę statystyczną danych. Obliczono szereg podstawowych parametrów opisowych, takich jak średnia arytmetyczna, odchylenie standardowe, kwartyle, rozstęp, wariancja, współczynnik zmienności oraz asymetrii. Wyniki wskazały, że średnia masa ciała wynosi około 86,6 kg, a dane wykazują lekką prawostronną asymetrię, co oznacza obecność osób z wyraźnie wyższą wagą niż przeciętna.

W kolejnym kroku przedstawiono cztery formy graficznej prezentacji danych: histogram, wykres pudełkowy, dystrybuantę empiryczną oraz wykres gęstości rozkładu. Wizualizacje potwierdziły wyniki analizy opisowej — rozkład danych jest lekko skośny w prawo, bez wyraźnych odstających wartości i z większością danych skoncentrowanych w przedziale 60–110 kg.

Następnie przeprowadzono dwa testy statystyczne:

- **Test średniej**, który wykazał, że masa ciała istotnie różni się od przyjętej wartości 70 kg (p-wartość $< 2{,}2 \times 10^{-16}$).

- **Test proporcji**, który potwierdził, że udział osób otyłych w badanej próbie jest znacząco wyższy niż zakładane 30% — faktycznie wynosi około 46%.

Oba testy pozwoliły na odrzucenie hipotez zerowych i wyciągnięcie statystycznie istotnych wniosków.

Wnioski końcowe:

1) Średnia masa ciała badanej grupy przekracza wartość referencyjną 70 kg, co może wskazywać na ogólny trend wzrostu masy ciała w populacjach latynoamerykańskich.

2) Udział osób otyłych jest znacznie wyższy niż zakładane 30%, co może wymagać szerszych działań profilaktycznych w zakresie zdrowia publicznego.

3) Analiza potwierdziła przydatność środowiska R w prowadzeniu badań statystycznych — umożliwia ono zarówno precyzyjne obliczenia, jak i atrakcyjne wizualizacje.

Zrealizowany projekt stanowi przykład praktycznego wykorzystania narzędzi statystycznych w analizie danych zdrowotnych, a uzyskane wyniki mogą stanowić podstawę do dalszych, pogłębionych badań w dziedzinie epidemiologii i zdrowia publicznego.

Załączniki

Listing 4.1. Kod użyty do analizy danych

```r
dane = read.csv("C:/Users/natal/Desktop/ObesityDataSet_raw_and_data_sinthetic.csv",
    sep = ",", header = TRUE)
waga = dane$Weight
srednia = mean(waga)
srednia
odchylenie = sd(waga)
odchylenie
kwantyle = quantile(waga)
kwantyle
wariancja = var(waga)
wariancja
wsp_zmiennosci =  odchylenie / srednia
wsp_zmiennosci
odchylenie_cwiartkowe =1/2*(kwantyle["75%"]- kwantyle["25%"])
odchylenie_cwiartkowe
rozstep = kwantyle["100%"]- kwantyle["0%"]
rozstep
library("e1071")
wsp_asymetrii = skewness(waga)
wsp_asymetrii
hist(waga,
     main = "Histogram masy ciaa",
     xlab = "Waga (kg)",
     col = "skyblue",
     border = "white")

plot(ecdf(waga),
     main = "Dystrybuanta empiryczna",
     xlab = "Waga (kg)",
     ylab = "Prawdopodobiestwo",
     col = "purple")

boxplot(waga,
        main = "Wykres pudekowy wagi",
        ylab = "Waga (kg)",
        col = "lightgreen")

plot(density(waga),
     main = "Wykres  gstoci  wagi",
     xlab = "Waga (kg)",
     col = "red")
# HIPOTEZA 1: Test redniej    czy  rednia  waga   rni    si  od 70 kg?
#  H :    = 70 kg (rednia waga wynosi 70)
#  H :       70 kg (rednia waga   rna   od 70)
t.test(waga, mu = 70)

# HIPOTEZA 2: Test proporcji  czy 30% osb to osoby otye?
#  H : p = 0.3 (30% populacji to osoby otye)
#  H : p    0.3

otyli <- dane$NObeyesdad %in% c("Obesity_Type_I", "Obesity_Type_II", "
    Obesity_Type_III")
x <- sum(otyli)
n <- length(otyli)
prop.test(x, n, p = 0.3)
```