Statystyczna analiza danych - projekt

Paweł Strzępka

Odczyt danych z pliku arkusza oraz ich obróbka

Dane pochodzą ze zbiorów Głównego Urzędu Statystycznego z Biutelynu Statystycznego Nr 4/2023 z dnia 25.05.2023 o częstotliwości miesięcznej. Zawierają dane z zakresu obszaru przemysłu gospodarki narodowej. Dotyczą produkcji sprzedanej przemysłu podanej w milionach PLN od roku 2010 do początku roku 2023.

```
getwd()
```

```
## [1] "D:/materialystudia/Statystycznaanaliza/Projekt"
```

```
library(readxl)
#### Odczyt  danych z pliku arkusza oraz ich obrobka ####
daneraw <- read_excel("tabl46_produkcja_sprzedana_przemyslu.xlsx")
dane <- daneraw[-c(1:3),]
dane <- dane[-2,]
dane0 <- dane[,1]
dane1 <- dane[,8]
dane2 <- dane[,10]
danew <- data.frame(dane0, dane1, dane2)
colnames(danew) <- c("Miesiac", "Przetworstwo przemyslowe", "Produkcja artykulow
         spozywczych")
danew <- danew[-1,]
danew[,2] <- as.numeric(danew[,2])
danew[,3] <- as.numeric(danew[,3])
danew
```

```
##      Miesiac Przetworstwo przemyslowe Produkcja artykulow spozywczych
## 2   2010 M01                  53435.6                          10545.7
## 3   2010 M02                  55489.7                          10532.7
## 4   2010 M03                  67369.8                          13023.2
## 5   2010 M04                  61939.4                          10900.3
## 6   2010 M05                  64897.9                          11814.8
## 7   2010 M06                  70275.9                          12120.8
## 8   2010 M07                  66276.0                          11849.2
## 9   2010 M08                  64774.8                          12166.0
## 10  2010 M09                  73338.2                          12550.8
## 11  2010 M10                  71518.1                          13035.0
## 12  2010 M11                  70224.0                          13037.8
## 13  2010 M12                  66536.8                          12931.0
## 14  2011 M01                  62625.6                          11997.4
## 15  2011 M02                  66101.8                          12263.5
## 16  2011 M03                  78737.6                          14066.8
## 17  2011 M04                  72349.6                          13458.1
## 18  2011 M05                  74670.9                          13255.1
## 19  2011 M06                  75837.3                          13418.4
## 20  2011 M07                  71648.5                          13152.5
## 21  2011 M08                  75819.4                          14392.9
## 22  2011 M09                  86207.2                          14927.2
## 23  2011 M10                  83699.1                          15465.6
## 24  2011 M11                  83850.0                          15607.4
## 25  2011 M12                  79116.9                          15342.3
## 26  2012 M01                  75052.1                          14266.5
## 27  2012 M02                  73870.6                          13944.0
## 28  2012 M03                  83108.7                          16081.4
## 29  2012 M04                  77773.9                          14489.5
## 30  2012 M05                  82768.5                          15592.0
## 31  2012 M06                  80474.5                          15069.3
## 32  2012 M07                  78587.1                          15126.4
## 33  2012 M08                  78131.4                          15670.6
## 34  2012 M09                  82961.6                          15566.2
## 35  2012 M10                  87885.5                          17356.2
## 36  2012 M11                  82854.9                          16593.8
## 37  2012 M12                  68851.4                          14946.5
## 38  2013 M01                  74140.0                          15654.1
## 39  2013 M02                  72397.6                          14444.3
## 40  2013 M03                  80202.9                          16638.0
## 41  2013 M04                  78416.8                          15290.4
## 42  2013 M05                  79201.6                          15534.1
## 43  2013 M06                  82520.1                          15316.9
## 44  2013 M07                  83925.9                          16263.8
## 45  2013 M08                  79598.0                          16110.7
## 46  2013 M09                  87688.4                          16510.8
## 47  2013 M10                  91454.5                          17616.8
## 48  2013 M11                  84584.1                          16650.7
```

```
## 49  2013 M12                74181.2                15844.7
## 50  2014 M01                77818.6                15681.4
## 51  2014 M02                76574.4                15068.4
## 52  2014 M03                85176.4                16123.1
## 53  2014 M04                83782.0                16484.7
## 54  2014 M05                82659.3                15628.3
## 55  2014 M06                82549.0                15533.8
## 56  2014 M07                84508.8                15552.1
## 57  2014 M08                76807.4                14992.8
## 58  2014 M09                90049.8                15752.2
## 59  2014 M10                92126.7                16636.7
## 60  2014 M11                83491.6                15430.5
## 61  2014 M12                78162.3                15588.0
## 62  2015 M01                77544.9                15008.5
## 63  2015 M02                79093.3                14442.9
## 64  2015 M03                90538.4                17073.9
## 65  2015 M04                83052.5                14669.2
## 66  2015 M05                82845.8                14863.9
## 67  2015 M06                88114.0                15500.3
## 68  2015 M07                86566.6                15680.8
## 69  2015 M08                78982.8                15552.4
## 70  2015 M09                91687.7                16327.0
## 71  2015 M10                92701.9                16917.6
## 72  2015 M11                89963.2                16470.6
## 73  2015 M12                84763.6                16277.3
## 74  2016 M01                78259.9                14929.6
## 75  2016 M02                84998.8                15310.0
## 76  2016 M03                91405.8                17193.0
## 77  2016 M04                89607.7                15868.8
## 78  2016 M05                87479.8                16299.2
## 79  2016 M06                94533.2                16774.2
## 80  2016 M07                84127.8                16357.6
## 81  2016 M08                86581.7                17276.9
## 82  2016 M09                96022.8                17641.0
## 83  2016 M10                93238.9                17971.2
## 84  2016 M11                95090.4                18350.8
## 85  2016 M12                90222.0                18239.2
## 86  2017 M01                89548.3                17167.4
## 87  2017 M02                89354.5                16584.2
## 88  2017 M03               106942.3                19608.0
## 89  2017 M04                92003.4                17526.7
## 90  2017 M05                97375.9                18411.3
## 91  2017 M06               100335.0                18290.7
## 92  2017 M07                91378.6                18028.1
## 93  2017 M08                96402.2                19486.7
## 94  2017 M09               104047.6                19112.8
## 95  2017 M10               108566.5                20388.8
## 96  2017 M11               107159.7                20104.5
```

```
## 97  2017 M12  93744.4  18529.3
## 98  2018 M01  98466.1  18488.4
## 99  2018 M02  95645.5  17599.0
## 100 2018 M03 107998.9  20671.4
## 101 2018 M04 101839.2  18259.8
## 102 2018 M05 105403.0  19247.1
## 103 2018 M06 110367.4  19351.2
## 104 2018 M07 104201.8  19192.0
## 105 2018 M08 104815.7  19784.3
## 106 2018 M09 109057.3  19219.8
## 107 2018 M10 120058.2  21550.3
## 108 2018 M11 114289.5  20926.0
## 109 2018 M12  97740.9  18797.0
## 110 2019 M01 105407.6  20030.3
## 111 2019 M02 104732.8  18806.3
## 112 2019 M03 117131.6  20805.0
## 113 2019 M04 114030.8  21113.9
## 114 2019 M05 114859.7  20433.3
## 115 2019 M06 106974.7  19148.9
## 116 2019 M07 111119.3  20438.4
## 117 2019 M08 103588.8  20024.9
## 118 2019 M09 115605.5  20479.1
## 119 2019 M10 123564.9  22248.9
## 120 2019 M11 113914.7  21279.2
## 121 2019 M12 101940.2  20570.8
## 122 2020 M01 107930.2  20909.8
## 123 2020 M02 110371.0  20232.0
## 124 2020 M03 112859.6  23383.9
## 125 2020 M04  81309.9  19145.5
## 126 2020 M05  91615.0  19431.4
## 127 2020 M06 106630.8  20644.8
## 128 2020 M07 110893.8  21381.5
## 129 2020 M08 103546.6  20480.7
## 130 2020 M09 121294.9  21669.1
## 131 2020 M10 124627.8  22463.4
## 132 2020 M11 121530.0  21455.5
## 133 2020 M12 114106.1  21337.8
## 134 2021 M01 109027.9  20113.5
## 135 2021 M02 115050.9  20822.1
## 136 2021 M03 140583.8  25843.3
## 137 2021 M04 128028.5  22071.5
## 138 2021 M05 128604.3  23058.5
## 139 2021 M06 135713.8  23582.4
## 140 2021 M07 131599.9  22838.6
## 141 2021 M08 129084.3  23721.8
## 142 2021 M09 144298.5  24860.8
## 143 2021 M10 146798.5  25383.9
## 144 2021 M11 154589.2  26254.8
```

```
## 145 2021 M12                146782.5                          26013.9
## 146 2022 M01                143201.1                          25123.0
## 147 2022 M02                151863.4                          26063.1
## 148 2022 M03                184993.9                          33648.6
## 149 2022 M04                171943.8                          31949.1
## 150 2022 M05                180369.4                          32416.2
## 151 2022 M06                181518.8                          32527.8
## 152 2022 M07                171266.6                          31790.5
## 153 2022 M08                172052.8                          33894.0
## 154 2022 M09                189096.2                          34720.5
## 155 2022 M10                188338.8                          35397.8
## 156 2022 M11                188900.6                          35687.8
## 157 2022 M12                173585.7                          34902.7
## 158 2023 M01                165250.5                          32661.9
```

Cecha nr 1 Przetworstwo przemyslowe

Wyznaczenie najmniejszej wartości

```
#### Minimum C1 ####
minimumc1 <- min(danew$`Przetworstwo przemyslowe`)
minimumc1
```

```
## [1] 53435.6
```

Wyznaczenie największej wartości

```
#### Maskimum C1####
maksimumc1 <- max(danew$`Przetworstwo przemyslowe`)
maksimumc1
```

```
## [1] 189096.2
```

Rozstęp jest najprostszą miarą rozproszenia (zmienności). Jest różnicą między wartością maksymalną a minimalną ze zbioru obserwacji. Pokazuje zatem jedynie jaki jest zakres obserwacji

```
#### Rozstep C1####
rozstepc1 <- maksimumc1 - minimumc1
rozstepc1
```

```
## [1] 135660.6
```

Wartość średnia pochodzi z sumowania poszczególnych wyników i podzielenie tej sumy przez liczbę naszych obserwacji.

```
#### Srednia C1####

sredniac1 <- mean(danew$`Przetworstwo przemyslowe`)
sredniac1
```

```
## [1] 99792.19
```

Mediana to wartość cechy w szeregu uporządkowanym, powyżej i poniżej której znajduje się jednakowa liczba obserwacji. Mediana jest kwantylem rzędu 1/2

```
#### Mediana C1####

medianac1 <- median(danew$`Przetworstwo przemyslowe`)
medianac1
```

```
## [1] 91378.6
```

Odchylenie standardowe określa, jak szeroko wartości jakiejś wielkości są rozrzucone wokół jej średniej

```
#### Odchylenie standardowe C1####

odchyleniec1 <- sd(danew$`Przetworstwo przemyslowe`)
odchyleniec1
```

```
## [1] 29589.14
```

Kwartyle to wartości, które dzielą zebrane obserwacje na cztery równe, co do ilości elementów, grupy.

```
#### Kwartyle C1####

kwartylec1 <- quantile(danew$`Przetworstwo przemyslowe`)
kwartylec1
```

```
##       0%      25%      50%      75%     100% 
##  53435.6  79201.6  91378.6 110371.0 189096.2
```

Wysoka wartość współczynnika oznacza duże zróżnicowanie cechy i świadczy o niejednorodności badanej populacji, niska wartość świadczy o małej zmienności cechy i jednorodności badanej populacji. Współczynnik zmienności jest ilorazem (wynikiem dzielenia) odchylenia standardowego cechy oraz jej średniej arytmetycznej.

```
#### Wspolczynnik zmiennosci C1####

wszmiennoscic1 <- (odchyleniec1/sredniac1) * 100
wszmiennoscic1
```

```
## [1] 29.65076
```

Wariancja informuje, jak bardzo zróżnicowany jest zbiór pod kątem koncentracji wokół średniej bądź też rozproszenia. Wartość zero oznacza identyczne wartości w zbiorze.

```
#### Wariancja C1####

wariancjac1 <- var(danew$`Przetworstwo przemyslowe`)
wariancjac1
```

```
## [1] 875517367
```

Trzeci moment centralny przyjmuje wartość zero dla rozkładu symetrycznego, wartości ujemne dla rozkładów o lewostronnej asymetrii i wartości dodatnie dla rozkładów o prawostronnej asymetrii.

```
#### Moment centralny rzedu 3 C1####

library(moments)
M3C1 <- moment(danew$`Przetworstwo przemyslowe`, order=3, central=TRUE)
M3C1
```

```
## [1] 3.496676e+13
```

Jeśli wartość współczynnika asymetrii jest równy zero, oznacza to, że rozkład jest symetryczny. Ujemne wartości wskazują na skośność w lewo, natomiast dodatnie wartości wskazują na skośność w prawo.

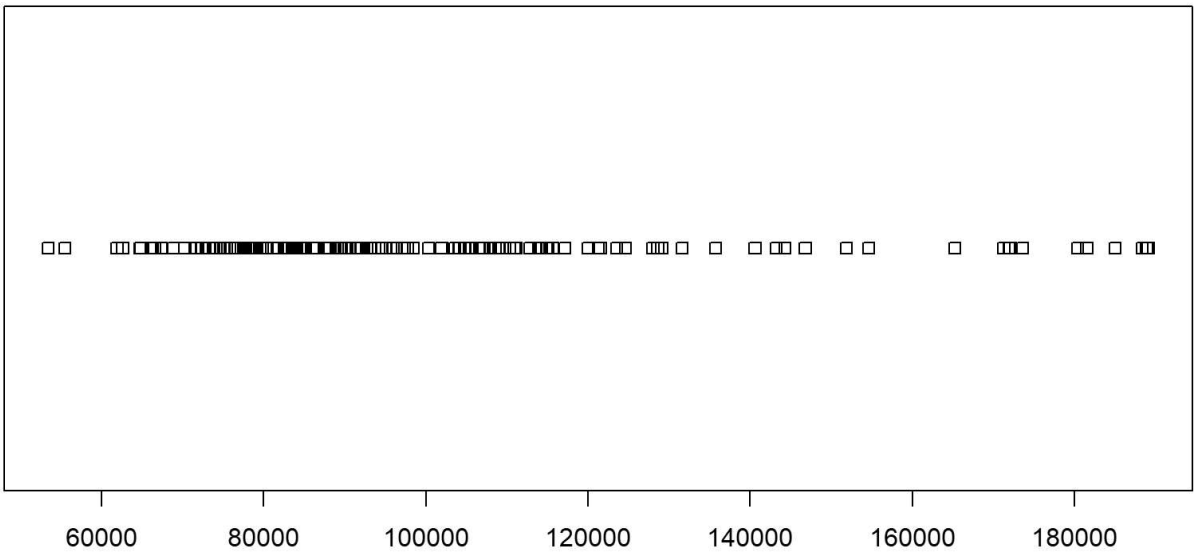
```
#### Wspolczynnik asymetrii ####

wsp_asymetriic1 <- skewness(danew$`Przetworstwo przemyslowe`)
wsp_asymetriic1
```

```
## [1] 1.362765
```

```
#### Graficzna reprezentacja danych C1####
```

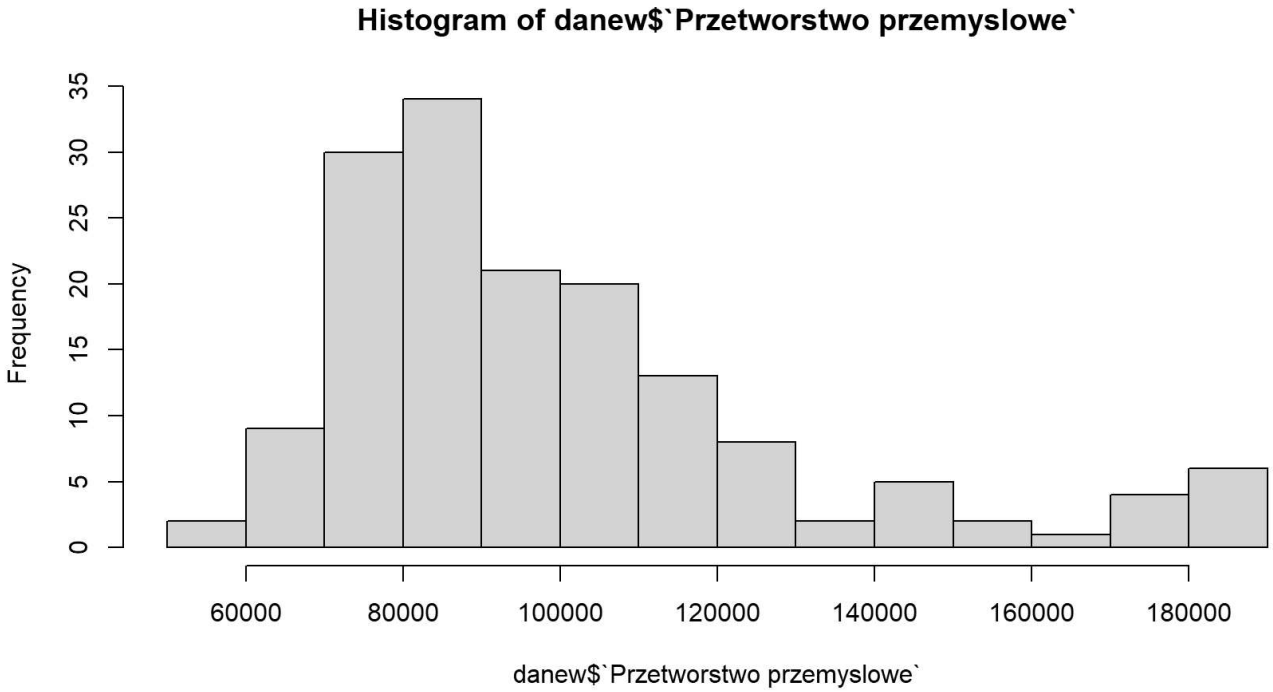
Wykres paskowy

```
#### Wykres paskowy ####
library(graphics)
stripchart(danew$`Przetworstwo przemyslowe`)
```



Histogram służy do przedstawienia liczebności obserwacji danych w zadanych przedziałach badanej zmiennej.

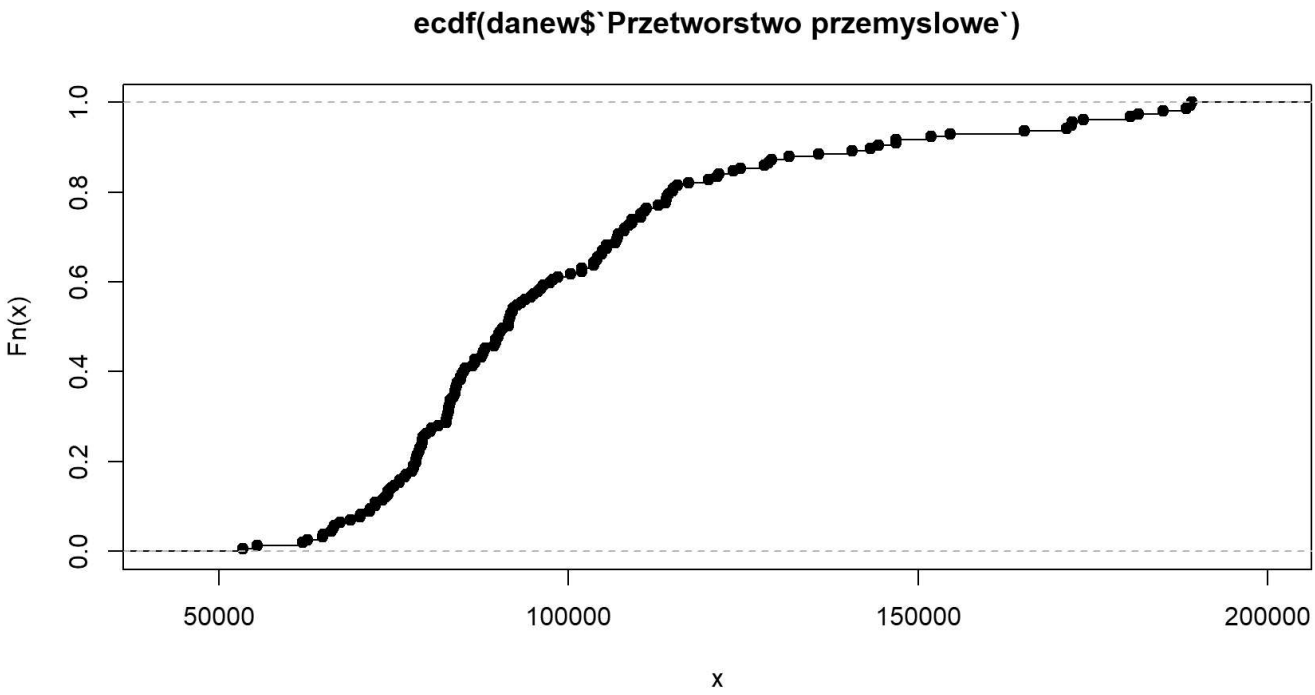
```
#### Histogram ####
(hist(danew$`Przetworstwo przemyslowe`, breaks=13))
```




```
## $breaks
##  [1]  50000  60000  70000  80000  90000 100000 110000 120000 130000 140000
## [11] 150000 160000 170000 180000 190000
##
## $counts
##  [1]  2  9 30 34 21 20 13  8  2  5  2  1  4  6
##
## $density
##  [1] 1.273885e-06 5.732484e-06 1.910828e-05 2.165605e-05 1.337580e-05
##  [6] 1.273885e-05 8.280255e-06 5.095541e-06 1.273885e-06 3.184713e-06
## [11] 1.273885e-06 6.369427e-07 2.547771e-06 3.821656e-06
##
## $mids
##  [1]  55000  65000  75000  85000  95000 105000 115000 125000 135000 145000
## [11] 155000 165000 175000 185000
##
## $xname
## [1] "danew$`Przetworstwo przemyslowe`"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

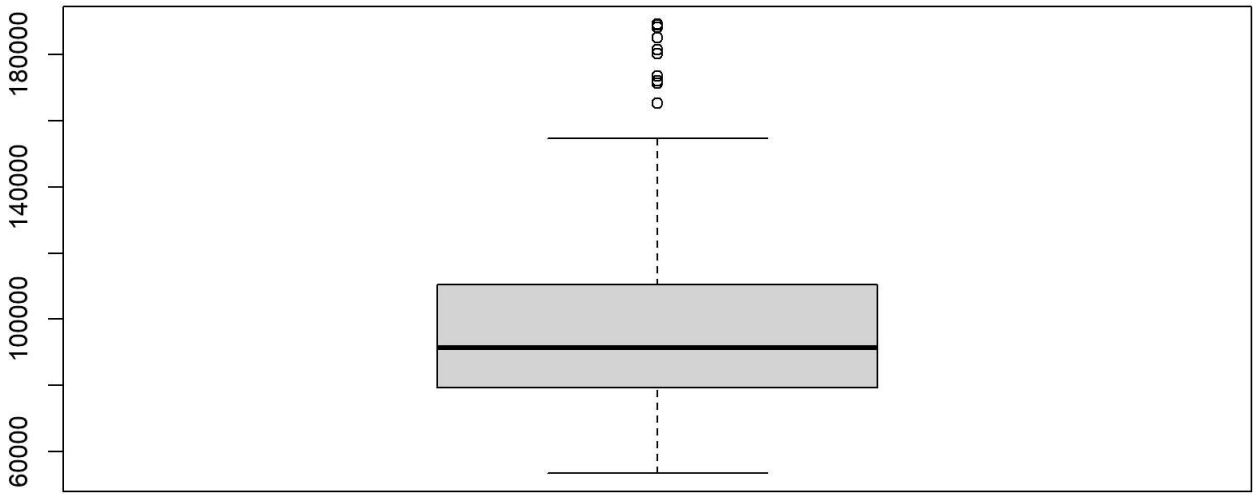
Dystrybuanta empiryczna to dystrybuanta wyliczona wprost z danych. W takiej sytuacji nie znamy prawdziwego rozkładu i bazujemy tylko na dostępnych obserwacjach

```
#### Dystrybuanta ####
plot(ecdf(danew$`Przetworstwo przemyslowe`))
```



Wykres pudełkowy zawiera informacje odnośnie położenia, rozproszenia i kształtu rozkładu danych. Zawiera mediane, rozstęp ćwiartkowy oraz wartości odstające, które odbiegają od reszty.

```
#### Wykres pudełkowy ####
boxplot(danew$`Przetworstwo przemyslowe`)
```



```
#### Hipotezy C1####
```

Hipoteza zerowa: Średnia wartość jest równa 100000 Hipoteza alternatywna: Średnia wartość nie jest równa 100000

Nie ma podstaw do odrzucenia hipotezy zerowej

```
#hipoteza 1 Średnia wartość jest równa 102000
t.test(danew$`Przetworstwo przemyslowe`,mu = 100000)
```

```
##
##  One Sample t-test
##
## data:  danew$`Przetworstwo przemyslowe`
## t = -0.088002, df = 156, p-value = 0.93
## alternative hypothesis: true mean is not equal to 1e+05
## 95 percent confidence interval:
##   95127.6 104456.8
## sample estimates:
## mean of x
##  99792.19
```

Hipoteza zerowa: Populacja ma rozkład normalny Hipoteza alternatywna: Populacja nie ma rozkładu normalnego

Odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej

```
# hipoteza 2 dane mają rozkład normalny
shapiro.test(danew$`Przetworstwo przemyslowe`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  danew$`Przetworstwo przemyslowe`
## W = 0.86872, p-value = 1.649e-10
```

Cecha nr 2 Produkcja artykulow spozywczych

```
#### Minimum C2####
minimumc2 <- min(danew$`Produkcja artykulow spozywczych`)
minimumc2
```

```
## [1] 10532.7
```

```
#### Maskimum C2####
maksimumc2 <- max(danew$`Produkcja artykulow spozywczych`)
maksimumc2
```

```
## [1] 35687.8
```

```
#### Rozstep C2####
rozstepc2 <- maksimumc2 - minimumc2
rozstepc2
```

```
## [1] 25155.1
```

```
#### Srednia C2####

sredniac2 <- mean(danew$`Produkcja artykulow spozywczych`)
sredniac2
```

```
## [1] 18646.42
```

```
#### Mediana C2####

medianac2 <- median(danew$`Produkcja artykulow spozywczych`)
medianac2
```

```
## [1] 17193
```

```
#### Odchylenie standardowe C2####

odchyleniec2 <- sd(danew$`Produkcja artykulow spozywczych`)
odchyleniec2
```

```
## [1] 5319.669
```

```
#### Kwantyle C2####

kwartylec2 <- quantile(danew$`Produkcja artykulow spozywczych`)
kwartylec2
```

```
##      0%     25%     50%     75%    100%
## 10532.7 15430.5 17193.0 20570.8 35687.8
```

```
#### Wspolczynnik zmiennosci C2####

wszmiennoscic2 <- (odchyleniec2/sredniac2) * 100
wszmiennoscic2
```

```
## [1] 28.52917
```

```
#### Wariancja C2####

wariancjac2 <- var(danew$`Produkcja artykulow spozywczych`)
wariancjac2
```

```
## [1] 28298875
```

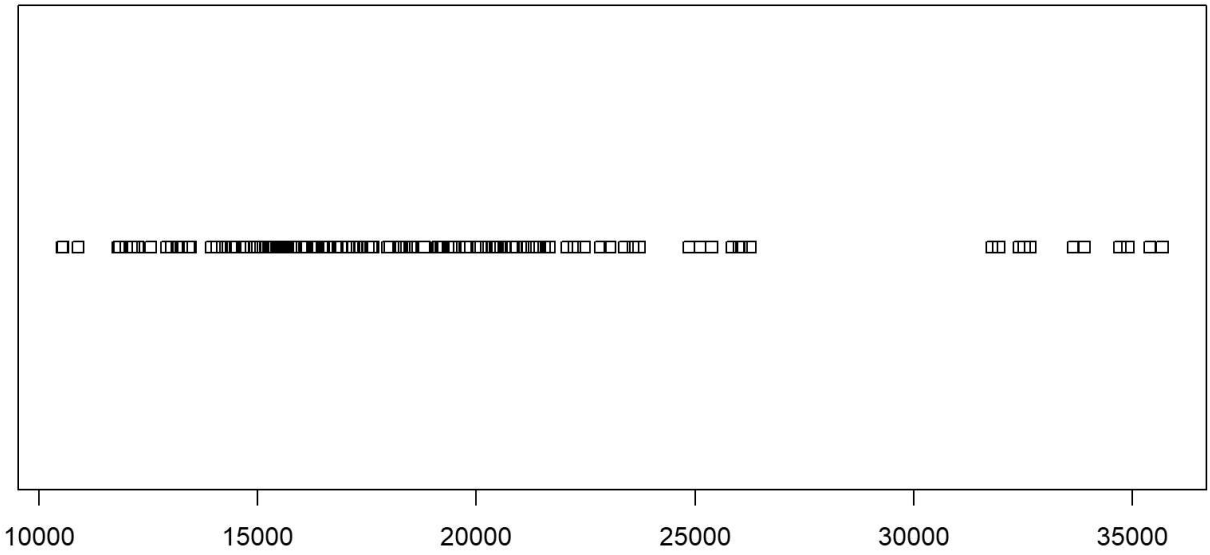
```
#### Moment centralny rzedu 3 C2####

library(moments)

M3C2 <- moment(danew$`Produkcja artykulow spozywczych`, order=3, central=TRUE)
M3C2
```

```
## [1] 218147285949
```

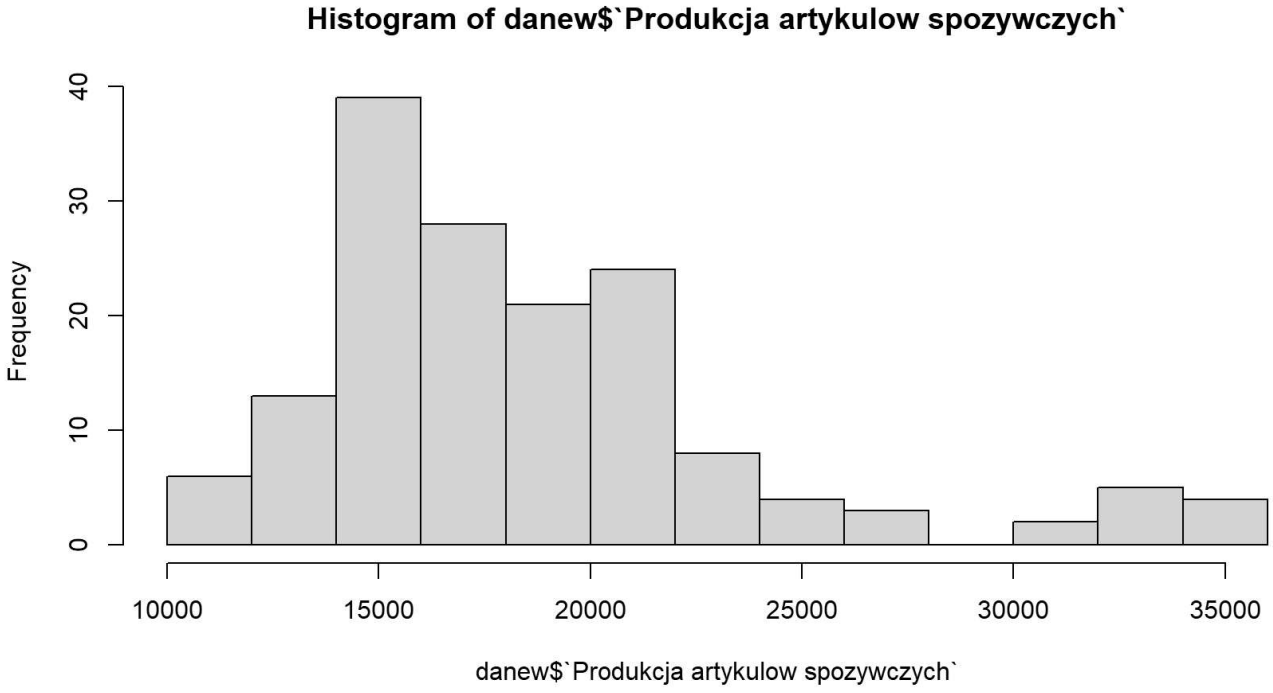
```
#### Wspolczynnik asymetrii ####

wsp_asymetriic2 <- skewness(danew$`Produkcja artykulow spozywczych`)
wsp_asymetriic2
```

```
## [1] 1.463048
```

```
#### Graficzna reprezentacja danych C2####


#### Wykres paskowy ####
library(graphics)
stripchart(danew$`Produkcja artykulow spozywczych`)
```

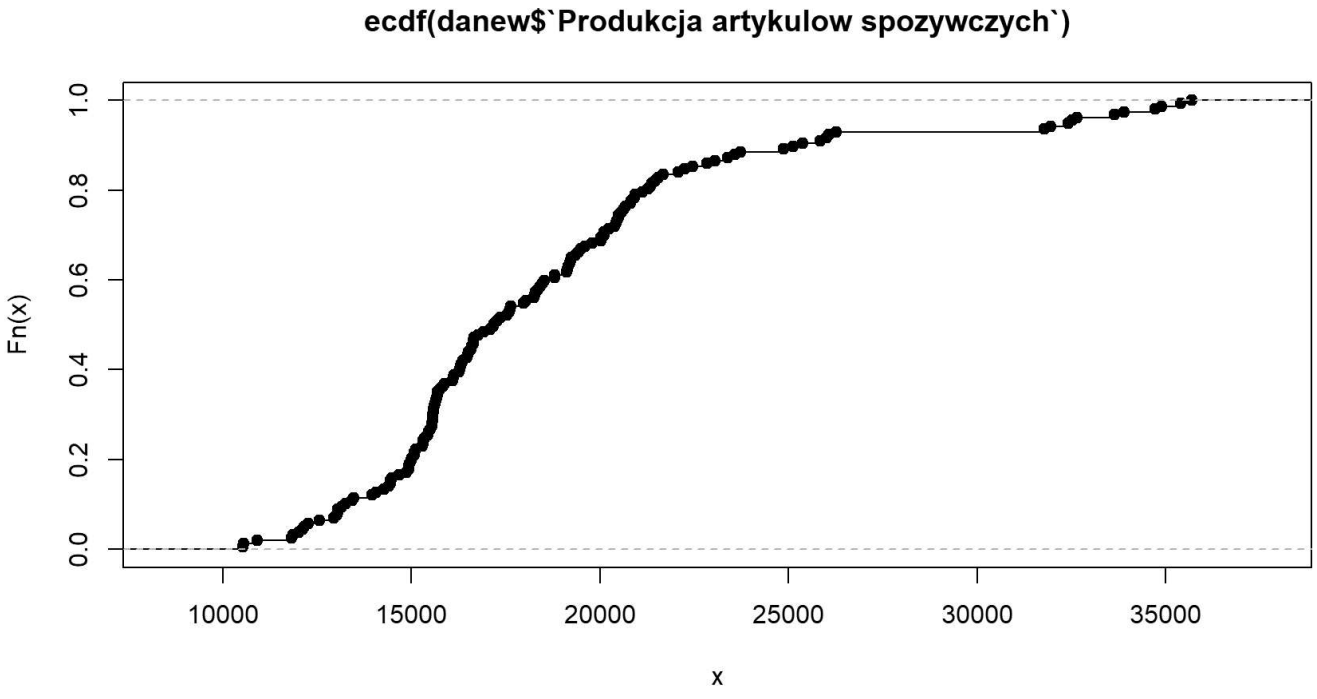



```
#### Histogram ####
(hist(danew$`Produkcja artykulow spozywczych`, breaks=13))
```

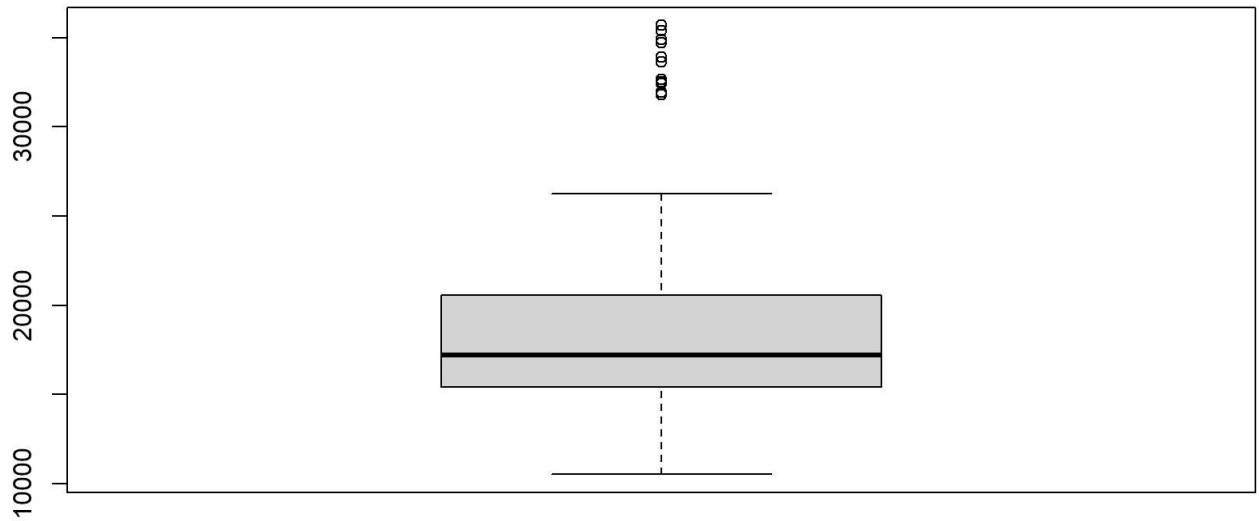


```
## $breaks
##  [1] 10000 12000 14000 16000 18000 20000 22000 24000 26000 28000 30000 32000
## [13] 34000 36000
##
## $counts
##  [1]  6 13 39 28 21 24  8  4  3  0  2  5  4
##
## $density
##  [1] 1.910828e-05 4.140127e-05 1.242038e-04 8.917197e-05 6.687898e-05
##  [6] 7.643312e-05 2.547771e-05 1.273885e-05 9.554140e-06 0.000000e+00
## [11] 6.369427e-06 1.592357e-05 1.273885e-05
##
## $mids
##  [1] 11000 13000 15000 17000 19000 21000 23000 25000 27000 29000 31000 33000
## [13] 35000
##
## $xname
## [1] "danew$`Produkcja artykulow spozywczych`"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
#### Dystrybuanta ####
plot(ecdf(danew$`Produkcja artykulow spozywczych`))
```



```
#### Wykres pudełkowy ####
boxplot(danew$`Produkcja artykulow spozywczych`)
```



Hipoteza zerowa: Średnia wartość jest równa 19000 Hipoteza alternatywna: Średnia wartość nie jest równa 19000

Nie ma podstaw do odrzucenia hipotezy zerowej

```
#### Hipotezy C2####

#hipoteza 1 Średnia wartość jest równa 19000
t.test(danew$`Produkcja artykulow spozywczych`,mu = 19000)
```

```
##
##  One Sample t-test
##
## data:  danew$`Produkcja artykulow spozywczych`
## t = -0.83283, df = 156, p-value = 0.4062
## alternative hypothesis: true mean is not equal to 19000
## 95 percent confidence interval:
##  17807.80 19485.04
## sample estimates:
## mean of x
##  18646.42
```

Hipoteza zerowa: Populacja ma rozkład normalny Hipoteza alternatywna: Populacja nie ma rozkładu normalnego

Odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej

```
# hipoteza 2 dane mają rozkład normalny
shapiro.test(danew$`Produkcja artykulow spozywczych`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  danew$`Produkcja artykulow spozywczych`
## W = 0.86382, p-value = 9.605e-11
```

Opis użytych funkcji:

getwd() - funkcja służy do pobrania bieżącego katalogu roboczego.

read_excel() - funkcja z pakietu “readxl” służy do odczytu pliku Excel do R i przechowywania danych w zmiennej.

data.frame() - funkcja z pakietu “base” tworzy ramkę danych na podstawie wektorów lub zmiennych. Używana jest do utworzenia nowej ramki danych o nazwie danew, łączącej wybrane kolumny z ramki danych daneraw.

colnames() - funkcja z pakietu “base” służy do nadania nazw kolumnom w ramce danych danew.

as.numeric() - funkcja z pakietu “base” służy do konwersji wybranych kolumn danew na format liczbowy.

min() - funkcja z pakietu “base” oblicza najmniejszą wartość wektora lub kolumny.

max() - funkcja z pakietu “base” oblicza największą wartość wektora lub kolumny.

mean() - funkcja z pakietu “base” oblicza średnią arytmetyczną wektora lub kolumny.

median() - funkcja z pakietu “stats” oblicza medianę wektora lub kolumny.

sd() - funkcja z pakietu “stats” oblicza odchylenie standardowe wektora lub kolumny.

quantile() - funkcja z pakietu “stats” oblicza kwantyle wektora lub kolumny.

var() - funkcja z pakietu “stats” oblicza wariancję wektora lub kolumny.

moment() - funkcja pochodzi z pakietu “moments”, który udostępnia funkcje do obliczania momentów rozkładu. Została użyta do obliczenia trzeciego momentu centralnego

skewness() - funkcja z pakietu “moments” oblicza współczynnik asymetrii dla danej populacji. W wyniku otrzymasz wartość współczynnika asymetrii.

stripchart() - funkcja z pakietu “graphics” pochodzi z podstawowego pakietu graficznego w R. Służy do tworzenia wykresu paskowego (diagram punktowy).

hist() - funkcja z pakietu “graphics” służy do tworzenia histogramu dla określonej zmiennej.

ecdf() - funkcja z pakietu “stats” służy do tworzenia empirycznej funkcji dystrybucji skumulowanej (ECDF) dla określonej zmiennej.

boxplot() - funkcja z pakietu “graphics” służy do tworzenia wykresu pudełkowego dla określonej zmiennej.

t.test() - funkcja z pakietu “stats” służy do przeprowadzania testu t, który porównuje średnią próbkową z daną wartością lub wykonuje test t dwóch próbek.

shapiro.test() - funkcja z pakietu “stats” służy do przeprowadzania testu Shapiro-Wilka, który testuje hipotezę zerową, że populacja ma rozkład normalnego.