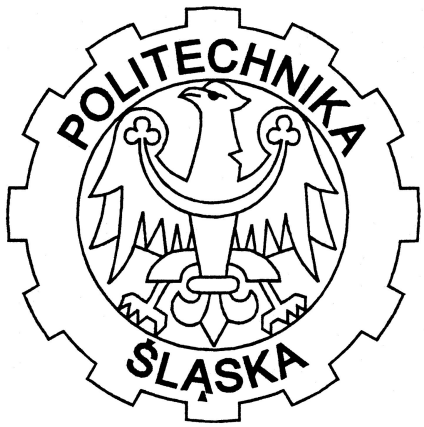
# Sprawozdanie

## Projekt z przedmiotu Metody Statystyczne

Temat ćwiczenia:
Projekt nr 8




AEiI, informatyka, sem. 4, 2015/2016
Grupa dziekańska: 3
Numer sekcji projektowej: 5

Skład sekcji:

Przemyslaw Pawlas
Radosław Wojaczek
Zbigniew Kmonk
Rafał Grzelec
Dawid Kubów

# Temat projektu

*“Wśród losowo wybranych klientów dóch marketów przeprowadzono badanie dotyczące miesięcznych wydatków na jedną osobę (w złotych), na pieczywo i produkty zbożowe. W pierwszym markecie uzyskano następujące odpowiedzi:*

56,2; 51,97; 38,63; 40,38; 36,56; 39,27; 60,56; 47,08; 46,51; 34,06; 45,36; 31,81; 39,95; 56,52; 51,27; 48,58; 29,61; 46,28; 43,87; 49,45; 33,38; 32,67; 51,61; 48,83; 43,73; 37,5; 52,54; 31,44; 38,6; 51,23; 55,65; 42,93; 54,69; 43,36; 21,22; 64,39; 31,99; 54,83; 51,95; 27,08; 36,35; 50,82

*W drugim markecie wyniki badań były następujące:*

34,92; 27,72; 28,31; 44,99; 39,63; 44,36; 46,45; 59,64; 32,8; 41,07; 44,17; 25,98; 40,04; 45,76; 43,53; 34,07; 38,23; 36,9; 40,9; 50,53; 55,31; 50,35; 64,78; 32,17; 45,46; 45,24; 28,92; 71,31; 39,75; 60,04; 65,15; 52,95; 21,14; 40,31; 60,93; 35,54; 47,05; 33,78; 54,16; 40,46; 47,86; 37,99; 31,18; 54,73; 63,11; 56,48; 36,1”

# Zadanie 1

*“Dokonać analizy miesięcznych wydatków na jedną osobę, na pieczywo i produkty zbożowe klientów wybranych marketów, wyznaczając miary przeciętne, zróżnicowania, asymetrii i koncentracji. Opracować histogramy rozkładów empirycznych. Miary wyznaczyć dwoma sposobami: a) na podstawie szeregu szczegółowego, b) na podstawie szeregu rozdzielczego.”*

Aby wyznaczyć wszystkie miary wyszczególnione w zadaniu, skorzystaliśmy z następujących wzorów:

| | Szereg szczegółowy | Szereg rozdzielczy |
|---|---|---|
| Średnia | $\frac{1}{n}\sum_{i=1}^{n} x_i$ | $\frac{1}{N}\sum_{i=1}^{k} \dot{x}_i n_i$ |
| Mediana | $\begin{cases} \frac{n}{2} \\ \frac{n+1}{2} \end{cases}$ | $\begin{cases} x_0 + (\frac{n}{2} - n_{isk-1})\frac{h_{Me}}{n_{Me}} \\ x_0 + (\frac{n+1}{2} - n_{isk-1})\frac{h_{Me}}{n_{Me}} \end{cases}$ |

| | | |
|---|---|---|
| Moda (dominanta) | | $x_{i_0} + \frac{n_{i_0} - n_{i_0-1}}{(n_{i_0} - n_{i_0-1}) + (n_{i_0} - n_{i_0+1})} \cdot h_{i_0}$ |
| Kwartyl pierwszy | $\begin{cases} \frac{n}{4} \\ \frac{n+1}{4} \end{cases}$ | $\begin{cases} x_0 + (\frac{n}{4} - n_{isk-1})\frac{h_{Me}}{n_{Me}} \\ x_0 + (\frac{n+1}{4} - n_{isk-1})\frac{h_{Me}}{n_{Me}} \end{cases}$ |
| Kwartyl trzeci | $\begin{cases} \frac{3n}{4} \\ \frac{3(n+1)}{4} \end{cases}$ | $\begin{cases} x_0 + (\frac{3n}{4} - n_{isk-1})\frac{h_{Me}}{n_{Me}} \\ x_0 + (\frac{3(n+1)}{4} - n_{isk-1})\frac{h_{Me}}{n_{Me}} \end{cases}$ |
| Wariancja obciążona | $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x_i})^2$ | $\frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 \cdot n_i$ |
| Wariancja nieobciążona | $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x_i})^2$ | $\frac{1}{n-1}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 \cdot n_i$ |
| Odchylenie standardowe obciążone | $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x_i})^2}$ | $\sqrt{\frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 \cdot n_i}$ |
| Odchylenie standardowe nieobciążone | $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x_i})^2}$ | $\sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 \cdot n_i}$ |
| Odchylenie ćwiartkowe | $\frac{Q_3 - Q_1}{2}$ | |
| Odchylenie przeciętne od średniej | $\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$ | $\frac{1}{n}\sum_{i=1}^{k}\lvert x_i - \bar{x}\rvert \cdot n_i$ |
| Odchylenie przeciętne od mediany | $\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - Me\rvert$ | $\frac{1}{n}\sum_{i=1}^{k}\lvert x_i - Me\rvert \cdot n_i$ |
| Rozstęp | $x_{max} - x_{min}$ | |
| Współczynnik zmienności | $\frac{s}{\bar{x}} \cdot 100\%$ | |

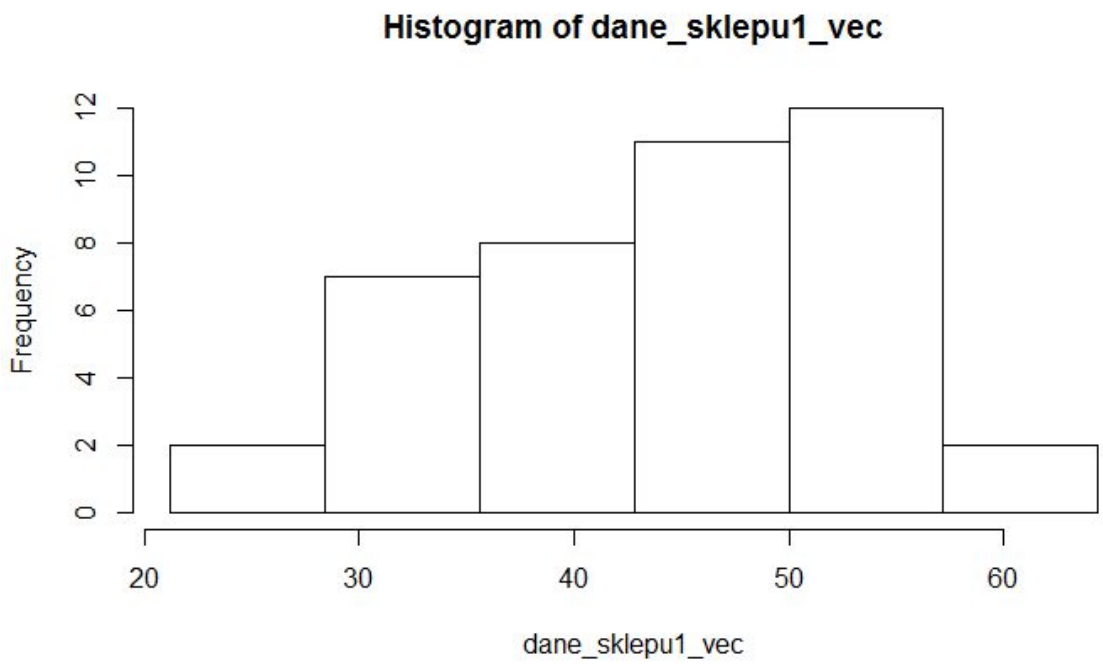
| | | |
|---|---|---|
| Współczynnik asymetrii | $\frac{1}{n}\sum_{i=1}^{n}[(x_i-\bar{x})^3]\cdot\frac{1}{s^3}$ | $\frac{1}{n}\sum_{i=1}^{k}[(x_i-\bar{x})^3\cdot n_i]\cdot\frac{1}{s^3}$ |
| Skosność | $\frac{3\cdot(\bar{x}-Me)}{s}$ | |
| Kurtoza | $\frac{1}{n}\sum_{i=1}^{n}[(x_i-\bar{x})^4]\cdot\frac{1}{s^4}$ | $\frac{1}{n}\sum_{i=1}^{k}[(x_i-\bar{x})^4\cdot n_i]\cdot\frac{1}{s^4}$ |
| Eksces | $K-3$ | |

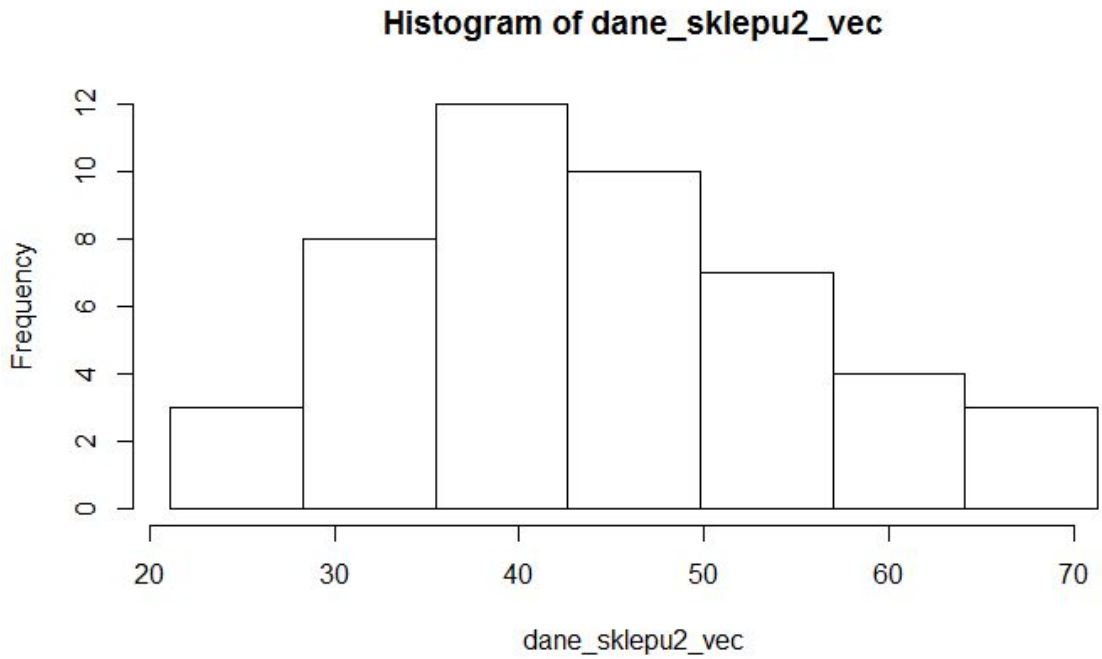
Po przeprowadzeniu translacji wzorów na język programowania R i uruchomieniu programu, otrzymaliśmy następujące wyniki:

| | Szereg szczegółowy | | Szereg rozdzielczy | |
|---|---|---|---|---|
| | Sklep pierwszy | Sklep drugi | Sklep pierwszy | Sklep drugi |
| Średnia | 44,065 | 44,09 | 44,347 | 44,243 |
| Mediana | 44,615 | 43,53 | 46,403 | 43,239 |
| Moda (dominanta) | Brak | Brak | 50,654 | 40,252 |
| Kwartyl pierwszy | 36,795 | 35,82 | 37,152 | 36,37 |
| Kwartyl trzeci | 51,525 | 51,74 | 52,289 | 51,959 |
| Wariancja obciążona | 93,791 | 130,508 | 86,984 | 128,315 |
| Wariancja nieobciążona | 96,078 | 133,345 | 89,106 | 131,105 |
| Odchylenie standardowe obciążone | 9,685 | 11,424 | 9,327 | 11,328 |
| Odchylenie standardowe nieobciążone | 9,802 | 11,548 | 9,44 | 11,45 |
| Odchylenie | 7,365 | 7,96 | 7,569 | 7,794 |

| | | | | |
|---|---|---|---|---|
| ćwiartkowe | | | | |
| Odchylenie przeciętne od średniej | 8,141 | 9,223 | 7,929 | 9,344 |
| Odchylenie przeciętne od mediany | 8,141 | 9,211 | 7,538 | 9,366 |
| Rozstęp | 43,17 | 50,17 | 43,17 | 50,17 |
| Współczynnik zmienności | 21,978% | 25,91% | 21,031% | 25,603% |
| Współczynnik asymetrii | -0,178 | 0,348 | -0,312 | 0,331 |
| Skosność | -0,171 | 0,147 | -0,661 | 0,266 |
| Kurtoza | 2,381 | 2,519 | 2,18 | 2,424 |
| Eksces | -0,619 | -0,481 | -0,82 | -0,576 |

Histogramy utworzone przez środowisko R:






**Wnioski:**

Porównując wyniki szeregów zauważamy, że pomimo tych samych danych wyniki niekoniecznie są identyczne, takie same uzyskaliśmy tylko w rozstępie, w innych miejscach wyniki różnią się już nieznacząco, lecz występują również większe różnice. Obliczenie i użycie naszych własnych punktów przerwań przedziałów w funkcji hist() zmniejszyło różnice pomiędzy niektórymi wartościami.

W przypadku mody (dominanty) zauważamy zasadniczą różnicę pomiędzy szeregami, mianowicie dla szeregu szczegółowego nie byliśmy w stanie uzyskać wyniku (wszystkie wartości występują tylko jeden raz), a dla szeregu rozdzielczego uzyskaliśmy.

# Zadanie 2

*"Sprawdzić, czy miesięczne wydatki na jedną osobę, na pieczywo i produkty zbożowe mają rozkład normalny (test zgodności Kołmogorowa-Lillieforsa, współczynik ufności 0,95)."*

Test zgodności Kołmogorowa – Lillieforse'a sprawdza czy rozkład w populacji dla pewnej zmiennej losowej różni się od założonego rozkładu teoretycznego, gdy znana jest pewna skończona liczba obserwacji tej zmiennej (próba statystyczna). Dodatkowo test ten często wykorzystywany jest w celu sprawdzenia czy zmienna ma rozkład normalny, co było naszym głównym celem.

Na podstawie danych pobranych z zadania wyznaczyliśmy dystrybuantę skumulowaną, wartość testową oraz różnicę między dystrybuantą skumulowaną a dystrybuantą dla rozkładu normalnego niezbędne do przeprowadzenia testu.

Następnie dla wyznaczonej maksymalnej różnicy między dystrybuantą skumulowaną a dystrybuantą dla rozkładu normalnego oraz wartości krytycznej sprawdziliśmy, czy podane rozkłady są normalne (tzn. czy wartość krytyczna jest większa od maksymalnej różnicy).
W obu przypadkach założenie to się sprawdza – oba testy podają, iż występuje rozkład normalny.

Wzór statystyki testowej:
$d = sup\,|F_n(x) - F_0(x)|$
Gdzie:
$F_n(x)$ - dystrybuanta empiryczna
$F_0(x)$ - dystrybuanta rozkładu normalnego

Postać zbioru krytycznego:
$K_0 = \ < k(1-\alpha),\ 1 >$
Gdzie:
$k(1-\alpha) = \frac{0.881}{\sqrt{n}}$

Wartość statystyki testowej dla zestawu danych sklepu 1: 0.06676955
Wartość statystyki testowej dla zestawu danych sklepu 2: 0.0936088
Wartość krytyczna dla zestawu danych sklepu 1: 0.1359413
Wartość krytyczna dla zestawu danych sklepu 2: 0.1285071

Wyniki z konsoli po uruchomieniu programu:

*"Wynik testu Kołmogorowa - Lillieforse'a dla zestawu danych sklepu 1*
*(H0 - podane dane mają rozkład normalny*
*H1 - nie mają…): brak podstaw do odrzucenia hipotezy zerowej - rozkład jest normalny."*
*"Wynik testu Kolmogorowa - Lillieforse'a dla zestawu danych sklepu 2*
*(H0 - podane dane mają rozkład normalny*
*H1 - nie mają): brak podstaw do odrzucenia hipotezy zerowej - rozkład jest normalny."*

**Wnioski:**

Udało nam się potwierdzić przypuszczenia, że wartości obu sklepów mają rozkład normalny. Pozwoliło nam to wykorzystać wzory dla rozkładu normalnego w kolejnych zadaniach, co znacznie uprościło pracę.

# Zadanie 3

*“Oszacować przedziałowo (współczynnik ufności 0,95) przeciętną wartość miesięcznych wydatków na jedną osobę, na pieczywo i produkty zbożowe klientów pierwszego marketu. Obliczyć względną precyzję oszacowania i sprawdzić, czy mamy podstawy do uogólniania otrzymanego przedziału ufności na całą populację miesięczynych wydatków, na jedną osobę, na pieczywo i produkty zbożowe klientów pierwszego marketu.”*

Używając współczynnika ufności 0.95 oraz danych sklepu 1 wraz z uzyskanymi wynikami z zadania 1 przystąpiliśmy do oszacowania przedziałów przeciętnej wartości.

Na poziomie istotności 0.05 i na podstawie liczebności danych możemy wyznaczyć przedziały:

a) gdy liczebność jest mała - n <= 30:

$$(\overline{x} - t(1-\tfrac{\alpha}{2}, n-1) * \tfrac{s}{\sqrt{n-1}},\ \overline{x} + t(1-\tfrac{\alpha}{2}, n-1) * \tfrac{s}{\sqrt{n-1}})$$

b) gdy liczebność jest duża - n > 30:

$$(\overline{x} - u(1-\tfrac{\alpha}{2}) * \tfrac{s}{\sqrt{n}},\ \overline{x} + u(1-\tfrac{\alpha}{2}) * \tfrac{s}{\sqrt{n}})$$

Następnie na podstawie błędu maksymalnego (różnicy górnej i dolnej wartości przedziału podzielonej przez 2) oraz średniej próby uzyskaliśmy precyzję względną.

Wyniki z konsoli po uruchomieniu programu:

*"Przedzial sredniej: ( 41.135629126748 , 46.9934184922996 )"*
*"Precyzja wzgledna: 6.646832% jest miedzy 5% a 10%, wiec możemy stwierdzic, ze istnieją podstawy do uogólnienia, jednak musimy pozostac ostrożni"*

**Wnioski:**

Precyzja względna pozwala nam uogólnić średnią próby, jednakże musimy być ostrożni.

# Zadanie 4

*“Oszacować przedziałowo (współczynnik ufności 0,95) odchylenie standardowe miesięcznych wydatków na jedną osobę, na pieczywo i produkty zbożowe klientów drugiego marketu. Obliczyć względną precyzję oszacowania i sprawdzić, czy mamy podstawy do uogólniania otrzymanego przedziału ufności na całą populację miesięczynych wydatków na jedną osobę, na pieczywo i produkty zbożowe klientów drugiego marketu.”*

Używając współczynnika ufności 0.95 oraz danych sklepu 2 wraz z uzyskanymi wynikami z zadania 1 przystąpiliśmy do oszacowania przedziałów odchylenia standardowego.

Na poziomie istotności 0.05 i na podstawie liczebności danych możemy wyznaczyć przedziały:

a) gdy liczebność jest mała - n <= 30:

$$\left(\sqrt{\frac{n * s^2}{\chi^2(1-\frac{\alpha}{2}, n-1)}},\ \sqrt{\frac{n * s^2}{\chi^2(\frac{\alpha}{2}, n-1)}}\right)$$

b) gdy liczebność jest duża - n > 30:

$$\left(\frac{s}{1+\frac{u(1-\frac{\alpha}{2})}{\sqrt{2n}}},\ \frac{s}{1-\frac{u(1-\frac{\alpha}{2})}{\sqrt{2n}}}\right)$$

Następnie na podstawie błędu maksymalnego (różnicy górnej i dolnej wartości przedziału podzielonej przez 2) oraz odchylenia standardowego obciążonego próby uzyskaliśmy precyzję względną.

Wyniki z konsoli po uruchomieniu programu:
*"Przedzial odchylenia: ( 9.5029334266443 , 14.3185648211095 )"*
*"Precyzja wzgledna: 21.076823% jest większe od 10%, wiec nalezy odrzucic teze, ze istnieją podstawy do uogólnienia"*

***Wnioski:***
Precyzja względna nie pozwala nam na uogólnienie.

# Zadanie 5

*"Czy na poziomie istotności 0,05 można twierdzić, że wartości miesięcznych wydatków na jedną osobę, na pieczywo i produkty zbożowe są większe dla klientów pierwszego marketu (sformułować i zweryfikować odpowiednią hipotezę)?"*

Po przeanalizowaniu treści zadania postanowiliśmy wysunąć następującą hipotezę:

Hipoteza $H_0$: $m_1 = m_2$
Kontrhipoteza $H_1$: $m_1 < m_2$

Na początku sprawdziliśmy równość wariancji w obu populacjach:

$H_0$ - są równe
$H_1$ - wariancja pierwszego sklepu jest mniejsza

Wykorzystujemy do tego test Fishera o statystyce:

$$F = \frac{wariancja\ nieobciążona\ pierwszego\ sklepu}{wariancja\ nieobciążona\ drugiego\ sklepu}$$

oraz obszarze krytycznym:

$$(-\infty, f[1-\alpha,\ m-1,\ n-1])$$

gdzie:
$\alpha$ - poziom istotności
m - liczebność danych pierwszego sklepu
n - liczebność danych drugiego sklepu
f - wartość odczytana funkcją qf()

Po podstawieniu wartości otrzymaliśmy:

F = 0.720526
Obszar krytyczny: $(-\infty,\ 1.650216)$

**Wniosek z testu**:
Wartość statystyki F zawiera się w obszarze krytycznym, wobec tego należy odrzucić hipotezę zerową. Wariancja populacji pierwszego sklepu jest mniejsza, stosujemy więc test Cochrana-Coxa aby porównać wartości przeciętne.

Gdy otrzymamy wynik testu, na jego podstawie należy sprawdzić wartości przeciętne:

a) gdy wariancje są równe wykonujemy test t-Studenta o statystyce:

$$C_n = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} \cdot \frac{m+n}{mn}}}$$

gdzie:

$\bar{x}$ - średnia próby
$s^2$ - wariancja próby

i obszarze krytycznym:
$(-\infty,\ t[1\ -\ \alpha,\ m\ +\ n\ -\ 2])$
gdzie:
t - wartość odczytana funkcją qt()

b) z kolei gdy wariancje są różne, przeprowadzamy test Cochrana-Coxa o statystyce:

$$C_n = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

i obszarze krytycznym:

$$\left(-\infty,\ \frac{\frac{S_1^2 * t(1-\alpha,\ m-1)}{m} + \frac{S_2^2 * t(1-\alpha,\ n-1)}{n}}{\frac{S_1^2}{m} + \frac{S_2^2}{n}}\right)$$

Na podstawie wyznaczonych wartości oraz tego czy zawierają się w przedziałach można stwierdzić czy nie ma podstaw do odrzucenia hipotezy zerowej $H_0$. W naszym wypadku wynikiem testu Cochrana-Coxa były:
Cn = -0.011572
Obszar krytyczny: (-∞, -1.680540)

**Wniosek z testu:**

Wartość statystyki C nie należy do obszaru krytycznego, zatem nie mamy podstaw do odrzucenia hipotezy zerowej. Możemy stwierdzić, że wartości przeciętne obu populacji są równe.

Wyniki z konsoli po uruchomieniu programu:

*Wynik testu: różne wariancje populacji [statystyka F = 0.720526, przedział (-∞, 1.650216), odrzucamy hipotezę zerową o równości], więc: statystyka C = -0.011572, obszar krytyczny (-∞, -1.680540). Wartość nie należy do przedziału - przyjmujemy hipotezę zerową - wartości przeciętne sa równe.*

**Wnioski do całego zadania:**

Test na danym poziomie istotności wykazał, że wartości przeciętne dla obu sklepów są równe, więc nasze założenie, że pierwszy sklep ma większą wartość przeciętną, było niepoprawne.

# Wnioski

Projekt zebrał całą wiedzę, którą nabyliśmy w trakcie ćwiczeń i wykładu. Zadania rozwiązane przez nas w projekcie wykorzystały wiele wzorów statystycznych, których działanie oraz zastosowanie mieliśmy okazję przetestować w trakcie realizacji celu.

Poznaliśmy nowy język programowania, którym jest R. Na początku sprawił nam on kilka trudności z powodu braku wiedzy na temat jego funkcjonowania, jednakże praca w środowisku sprawiła, że nauczyliśmy się nim posługiwać i zrobienie zadań nie stanowiło już większych przeszkód.

Wnioski do wyników poszczególnych zadań zostały zamieszczone powyżej.