01.07.2016 Gliwice

Metody Statystyczne

Sprawozdanie z projektu 7

Informatyka,
Sem. 4, Gr. 3
Skład sekcji:
Karol Tytko
Tomasz Matloch
Wojciech Kaziród
Marek Makowski
Paweł Rugor

Treść Projektu 7

Wśród losowo wybranych klientów dwóch marketów przeprowadzono badanie dotyczące miesięcznych wydatków na jedną osobę(w złotych), na pieczywo i produkty zbożowe.

W pierwszym markecie uzyskano następujące odpowiedzi:
56.20, 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, 37.50, 52.54, 31.44, 38.60, 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, 36.90, 40.90, 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, 3.78, 54.16, 40.46, 47.86, 37.99, 31.18, 54.73, 63.11, 56.48, 36.10

Polecenia do wykonania:

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.

Miary zostały wyznaczone na podstawie zamieszczonych wzorów:

a) Dla szeregu szczegółowego:
   - Miary przeciętne:

- średnia arytmetyczna: $\overline{x}=\sum_{i=1}^{n} x_i$

- średnia harmoniczna: $\overline{x_h}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$

- średnia geometryczna: $\overline{x_g}=\sqrt[n]{\prod_{i=0}^{n} x_i}$

- mediana $m_e=Q_2$

- Miary zróżnicowania:

- wariancja nieobciążona: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$

- odchylenie standardowe: $S = \sqrt{S^2}$

- odchylenie przeciętne od mediany: $d_2 = \frac{1}{n}\sum_{i=1}^{n}|x_i - m_e|$

- odchylenie ćwiartkowe: $Q = \frac{Q_1 - Q_3}{2}$

- współczynnik zmienności: $V = \frac{s}{\overline{x}}$

- Miary asymetrii rozkładu:

- współczynnik asymetrii: $A_s = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{s^3}$

- Miary koncentracji wartości w próbie:

- kurtoza: $Krt = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{s^4}$

b) dla szeregu rozdzielczego

- Miary przeciętne:

- wariancja: $S^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_{oi} - \overline{x}\right)^2 n_i = \frac{1}{n}\sum_{i=1}^{k} x_{oi}^2 n_i - \overline{x}^2$

- odchylenie standardowe: $S = \sqrt{S^2}$

- odchylenie przeciętne od mediany: $d_2 = \frac{1}{n}\sum_{i=1}^{k} |x_i - m_e| n_i$

- odchylenie ćwiartkowe: $Q = \frac{Q_1 - Q_3}{2}$

- współczynnik zmienności: $V = \frac{s}{\overline{x}}$

- Miary zróżnicowania

- średnia arytmetyczna: $\overline{x} = \sum_{i=1}^{k} x_{oi} n_i$

- średnia harmoniczna: $\overline{x_h} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

- średnia geometryczna: $\overline{x_g} = \sqrt[n]{\prod_{i=0}^{k} x_i^{n_i}}$

- mediana: $m_e = X_{lm} + \frac{b}{n_m}\left(\frac{n}{2} - \sum_{i=1}^{m-1} n_i\right)$

- Miary asymetrii rozkładu:

- kurtoza: $Krt = \frac{\frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^4}{s^4}$

- Miary koncentracji wartości w próbie:

- współczynnik asymetrii: $A_s = \frac{\frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^3}{s^3}$

Otrzymane wyniki:

Dane1:

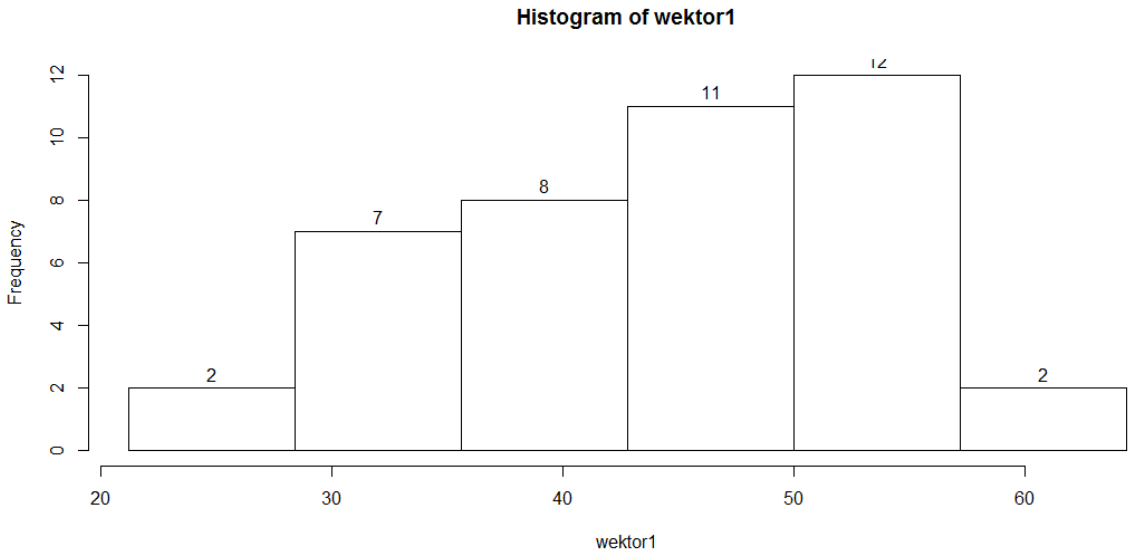
	szczegolowy	rozdzielczy
Srednia arytmetyczna:	44.0645238	44.0476190
Srednia harmoniczna:	41.6599533	41.6224126
Srednia geometryczna:	42.9123264	42.8721573
Wariancja nieobciazona:	96.0784644	97.0096372
Odchylenie standardowe:	9.8019623	9.8493470
Mediana:	44.6150000	45.0000000
Odchylenie przecietne od mediany:	8.1411905	8.4523810
Kwartyl Q1:	36.7950000	36.0714286
Kwartyl Q3:	51.5250000	51.9444444
Odchylenie cwiartkowe:	7.3650000	7.9365079
Wspolczynnik zmiennosci:	22.244566 %	22.3606798 %
Wspolczynnik asymetrii:	-0.1714869	-0.1170580
Wspolczynnik skosnosci:	-0.1684794	-0.2900845
Kurtoza:	2.2693594	2.1812719
Exces:	-0.7306406	-0.8187281

Dane2:

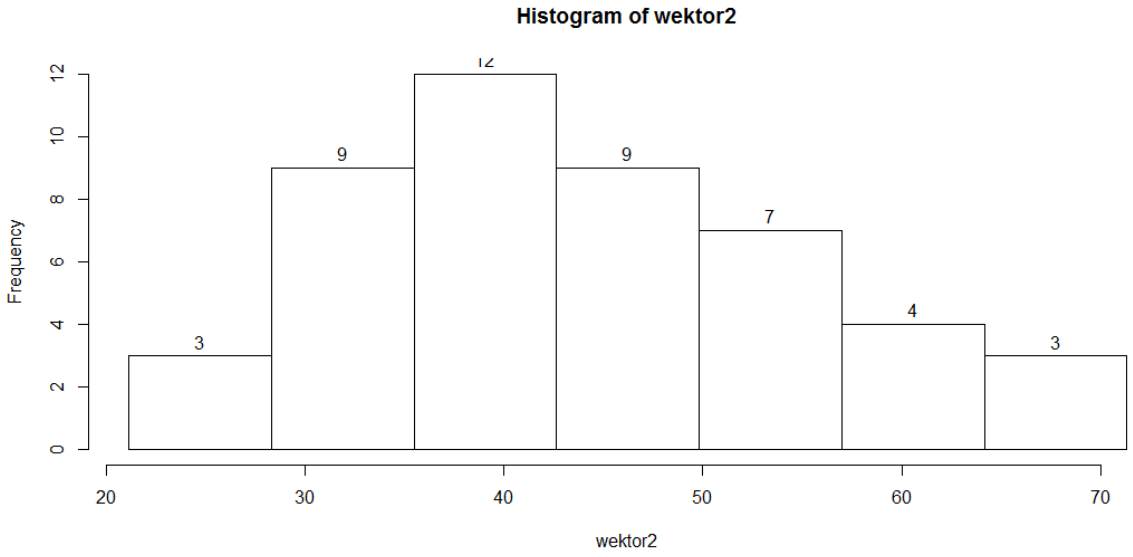
	szczegolowy	rozdzielczy
Srednia arytmetyczna:	43.8563830	44.3617021
Srednia harmoniczna:	40.8387308	40.9828180
Srednia geometryczna:	42.3470625	42.6666412
Wariancja nieobciazona:	135.4890801	150.6564056
Odchylenie standardowe:	11.6399777	12.2742171
Mediana:	41.0700000	43.2142857
Odchylenie przecietne od miediany:	9.3302128	9.9164134
Kwartyl Q1:	35.2300000	34.8214286
Kwartyl Q3:	51.7400000	52.8125000
Odchylenie cwiartkowe:	8.2550000	8.9955357
Wspolczynnik zmiennosci:	26.541125 %	27.668499 %
Wspolczynnik asymetrii:	0.3737222	0.3977740
Wspolczynnik skosnosci:	0.7181413	0.2804455
Kurtoza:	2.3766823	2.5318820
Exces:	-0.6233177	-0.4681180

Histogramy rozkładów empirycznych:

Dane z pierwszego sklepu:



Dane z drugiego sklepu:



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

Tablica testu Kołmogorowa-Lillieforsa

	poziom α	
n	0,01	0,05
31	0,1852	0,1591
32	0,1823	0,1566
33	0,1795	0,1542
34	0,1768	0,1519
35	0,1743	0,1498
36	0,1717	0,1477
37	0,1695	0,1457
38	0,1673	0,1437
39	0,1651	0,1419
40	0,1630	0,1401
41	0,1610	0,1384
42	0,1591	0,1367
43	0,1572	0,1351
44	0,1554	0,1336
45	0,1537	0,1321
46	0,1520	0,1306
47	0,1504	0,1292
48	0,1488	0,1279
49	0,1473	0,1266
50	0,1458	0,1253
51	0,1444	0,1241
52	0,1430	0,1229
53	0,1416	0,1217
54	0,1403	0,1206
55	0,1390	0,1193
60	0,1331	0,1144
65	0,1279	0,1099
70	0,1232	0,1059
75	0,1190	0,1023
80	0,1153	0,0991
85	0,1118	0,0961
90	0,1087	0,0934
95	0,1058	0,0909
100	0,1031	0,0886

H_0 - hipoteza zerowa, oznaczająca że wytrzymałości mają rozkład normalny

H_1 - hipoteza pierwsza, oznaczająca że wytrzymałości nie mają rozkładu normalnego

$$d_n = max(d_n^+ , d_n^-)$$
$$d_n^+ = \max_{1 \leqslant i \leqslant n} |\frac{i}{n} - F_0(x)|$$
$$d_n^- = \max_{1 \leqslant i \leqslant n} |F_0(x) - \frac{i-1}{n}|$$
$$K_0 = \langle d_n(1-\alpha), 1 \rangle$$

Gdzie:

$F_0(x)$ - jest funkcją rozkładu normalnego

$d_n(1-\alpha)$ - jest wartością odczytaną z tablicy

Patrzymy na 0.05, ponieważ w zadaniu jest współczynnik ufności 0,95.
1-0,95 = 0,05

```
> zadanie2(dane1$V1);
Wartość k:  0.1367
wartosc d:  0.08798346
Wydatki na jedna osobe na pieczywo i produkty zbozowe maja rozklad normalny.
> zadanie2(dane2$V1);
Wartość k:  0.1292
wartosc d:  0.105233
Wydatki na jedna osobe na pieczywo i produkty zbozowe maja rozklad normalny.
```

3. Czy na poziomie istotności 0,05 można twierdzić, że przeciętna wartość miesięcznych wydatków na jedną osobę, na pieczywo i produkty zbożowe dla klientów pierwszego marketu jest równa 42zł?

N = 42 – ilość danych z pierwszego marketu
α = 0,05 – poziom istotności

Z rozwiązania zadania 2 wiemy, że cecha ma rozkład normalny, więc korzystamy z rozkładu T-Studenta

Hipoteza zerowa

$H_0: \mu = \mu_0$

$\sigma = \sigma_0 = 42$

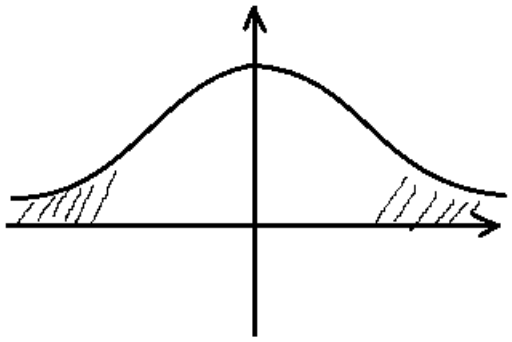
Następnie obliczamy wartość statystyki ze wzoru:

$$t = \frac{\overline{x} - \mu_0}{S} \sqrt{n-1}$$

Gdzie:
$\bar{x}$ – średnia z wektora (funkcja MEAN)
S – odchylenie standardowe z wektora (funkcja SD)

Przechodzimy do hipotezy alternatywnej

$H_1: \mu \neq \mu_0$



Obszar krytyczny:

$$K = \left((-\infty \; ; \; -t\left(1 - \frac{\alpha}{2}, n-1\right) \right) \cup \left(t\left(1 - \frac{\alpha}{2}, n-1\right) \; ; \; +\infty \;) \right)$$

Kwantyl t zostaje zwrócony przez użycie funkcji QT
Jeżeli $t \in K$ to nie mamy podstaw do odrzucenia hipotezy zerowej.

Funkcja którą napisaliśmy do obliczenia hipotezy przyjmuje 2 wartości:
-wektor
-alfa

```
> zadanie3(dane1$V1, 0.05)
Poziom istotnosci testu:  0.05
Hipoteza zerowa: przecietna == 42
Hipoteza alternatywna: przecietna != 42
Wartosc statystyki: 1.348649
Przedzialy krytyczne: ( -oo,  -2.019541 ) u  ( 2.019541 ,oo)
Brak podstaw do odrzucenia hipotezy zerowej.
```

4. Czy na poziomie istotności 0,05 można twierdzić, że odchylenie standardowe miesięcznych wydatków na jedną osobę, na pieczywo i produkty zbożowe dla klientów drugiego marketu jest równe 10zł?

N = 47 – ilość danych z pierwszego marketu
α = 0,05 – poziom istotności
σ = 10 - odchylenie standardowe

Zadanie rozwiązujemy podobnie jak poprzednie, lecz tutaj przyjmiemy statystykę Chi-kwadrat. Liczymy ją ze wzoru:

$$X^2 = \frac{nS^2}{\sigma_0^2}$$

gdzie:
n - ilość próbek
S^2 - wariancja

Hipoteza zerowa

$$H_0: \sigma = \sigma_0$$

$\sigma = \sigma_0 = 10$

Przechodzimy do hipotezy alternatywnej

$$H_1: \sigma \neq \sigma_0$$

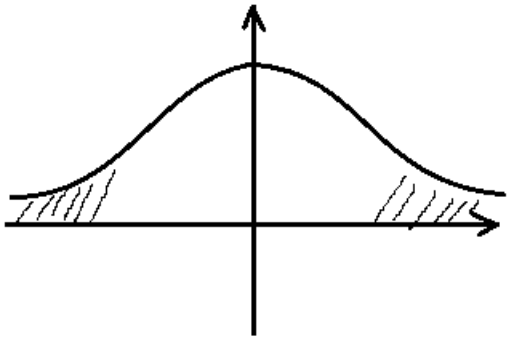
1) $\sigma \neq 10$

Następnie odczytujemy kwantyle z tablic.

Lewostronny dla wartości ($\frac{\alpha^2}{2}$, n−1)

Prawostronny dla wartości (1− $\frac{\alpha^2}{2}$, n−1),

gdzie (n - 1) to stopnie swobody.



Jeżeli nasza obliczona statystyka Chi-kwadrat, nie znajduje się w obszarach krytycznych, czyli jest większa od kwantyla lewostronnego, oraz mniejsza od kwantyla prawostronnego, to nie możemy odrzucić hipotezy, a odchylenie jest równe 10 zł. W przeciwnym wypadku odrzucamy hipotezę zerową na rzecz alternatywnej, odchylenie jest różne od 10 zł.

Funkcja którą napisaliśmy do obliczenia hipotezy przyjmuje 2 wartości:
-wektor
-alfa

```
> zadanie4(dane2$V1,10,0.05)
Poziom istotnosci testu:  0.05
Odchylenie standardowe:  10
Hipoteza alternatywna: odchylenie standardowe nie jest rowne  10
Wartosc statystyki: 63.67987
Przedzialy krytyczne: ( -oo,  -29.16005 ) u  ( 66.61653 ,oo)
Na poziomie istotnosci  0.05 nie mozemy odrzucic hipotezy,ze odchylenie standardowe miesieczny
ch wydatkow na jedna osobe, na pieczywo i produkty zbozowe dla klientow drugiego marketu jest
rowne 10
```

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)?

Na początku gdy porównujemy dwie średnie z prób przeprowadzonych na różnych populacjach, musimy dobrać odpowiednią statystykę. Aby dowiedzieć się jakiej statystyki użyć, należy przeprowadzić test Fishera-Snedecora porównujący wariancję prób.

Testujemy hipotezę zerową:

$$H_0: \sigma_X^2 = \sigma_Y^2$$

Przeciwko jednej z 3 wariantów hipotezy alternatywnej. Rozwiązując zadanie wybraliśmy następującą hipotezę alternatywną:

$$H_1: \sigma_X^2 > \sigma_Y^2$$

Czyli wariancja z próby pierwszej jest większą od wariancji z próby drugiej.
Użyliśmy następującej statystyki testowej:

$$F = \frac{S_X^2}{S_Y^2}$$

Gdy hipoteza H_0 jest prawdziwa to statystyka F ma rozkład F-Snedecora z (n-1) i (m-1) stopniami swobody. Należy również pamiętać przy obliczeniach, że w liczniku powinna się znaleźć wariancja o większej wartości.

Wyznaczamy kwantyl:

$$\left(f_{n-1,m-1}(1-\alpha),\ \infty\right))$$

```
> zadanie5(dane1$V1, dane2$V1, 0.05);
Test fishera-snedecora
Hipoteza zerowa: wariancje sa sobie rowne
Hipoteza alternatywna: wariancja z marketu pierwszego jest wieksza od wariancji z marketu drugigo
Wartosc statystyki fishera: 1.410192
Przedzial krytyczny: ( 1.664267 ,oo)
Nie ma podstaw do odrzucenia hipotezy zerowej, przyjmujemy ze wariancje sa w przyblizeniu rowne
```

Odpowiada on wybranej przez nas hipotezie alternatywnej.

Jeżeli wartość statystyki należy do obszaru krytycznego to uznajemy, że wariancje różnią się w sposób znaczący i w dalszej części zadania używamy statystyki Cochrana-Coxa. Jeżeli statystyka nie należy do obszaru krytycznego to używamy statystyki T-studenta.

Przechodzimy do kolejnej części zadania.
Jako hipotezę zerową przyjmujemy, że średnie z dwóch prób są równe.

$$H_0: m_1 = m_2$$

Hipoteza alternatywna będzie taka sama jak w poprzedniej części zadania, czyli przyjmujemy:

$$H_1: m_1 > m_2$$

Czyli wartość miesięcznych wydatków na jedną osobę na pieczywo i produkty zbożowe jest większa od wartości miesięcznych wydatków w drugim sklepie.
Jeżeli statystyka należy do obszaru krytycznego to możemy stwierdzić prawdziwość hipotezy alternatywnej, jeżeli statystyka nie należy do obszaru krytycznego, to możemy stwierdzić prawdziwość hipotezy zerowej. Obszar krytyczny będzie prawostronny.

Testowanie hipotez wykonujemy na takiej samej zasadzie jak w zadaniach 3 i 4.
Wykorzystujemy odpowiednie statystyki:

Statystyka **Cochrana-Coxa**

Statystyka ta wygląda następująco:

$$C = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

$\bar{X}_n$ i $\bar{Y}_m$ - odpowiednie średnie próbkowe
S_X^2 i S_Y^2 - odpowiednie wariancje próbkowe

Dla zadanych liczebności prób n i m można znaleźć przybliżoną wartość $C_{n,m}(1-\alpha)$ kwantyla $(1-\alpha)$ rzędu rozkładu statystyki C, a mianowicie

$$C_{n,m}(1-\alpha) \approx \frac{\frac{S_1^2}{n} t_{n-1}(1-\alpha) + \frac{S_2^2}{m} t_{m-1}(1-\alpha)}{\frac{S_1^2}{n} + \frac{S_2^2}{m}}$$

Obszar krytyczny jest prawostronny.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_1-1)s_1^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Statystyka **T-Studenta**

gdzie:
$\bar{x}_1$, $\bar{x}_2$ - średnie z prób
s_1^2, s_2^2 - odchylenia standardowe prób
n_1, n_2 - liczebności prób

Rozkład statystyki testowej t jest rozkładem T-Studenta o n_1+n_2-2 stopniach swobody.

Otrzymano następujące wyniki:

```
> zadanie5(dane1$V1, dane2$V1, 0.05);
Test fishera-snedecora
Hipoteza zerowa: wariancje sa sobie rowne
Hipoteza alternatywna: wariancja z marketu pierwszego jest wieksza od wariancji z marketu drugigo
Wartosc statystyki fishera: 1.410192
Przedzial krytyczny: ( 1.664267 ,oo)
Nie ma podstaw do odrzucenia hipotezy zerowej, przyjmujemy ze wariancje sa w przyblizeniu rowne
Statystyka t studenta
hipoteza zerowa: srednie sa rowne
Hipoteza alternatywna: srednia wartosc miesiecznych wydatkow w pierwszym markecie jest wieksza od sredniej wartosci miesiec
znych wydatkow w drugim markecie
Wartosc statystyki T-studenta: 0.08964174
Wartosc graniczna przedzialu krytycznego(najwyzsza): 1.662557
Na poziomie istotnosci  0.05 nie mozna przyjac hipotezy ze wartosc miesiecznych wydatkow na osobe w pierwszym markecie jest
 wieksza
```

Ostatnie zadanie w projekcie, wymagało od nas większej niż standardowej wiedzy o zagadnieniach z zakresu statystyki. Wymagane było tu przeprowadzenie 3 różnych testów(sprawdzenie czy dane mają rozkład normalny nastąpiło w zadaniu 2). Tego typu zadania, mogą być już przydatne w badaniach np. rynkowych, więc tego typu zadanie może nam się w przyszłości przydać.