Sprawozdanie

Projekt z przedmiotu Metody Statystyczne

Temat ćwiczenia:

Projekt nr 8

Skład sekcji:

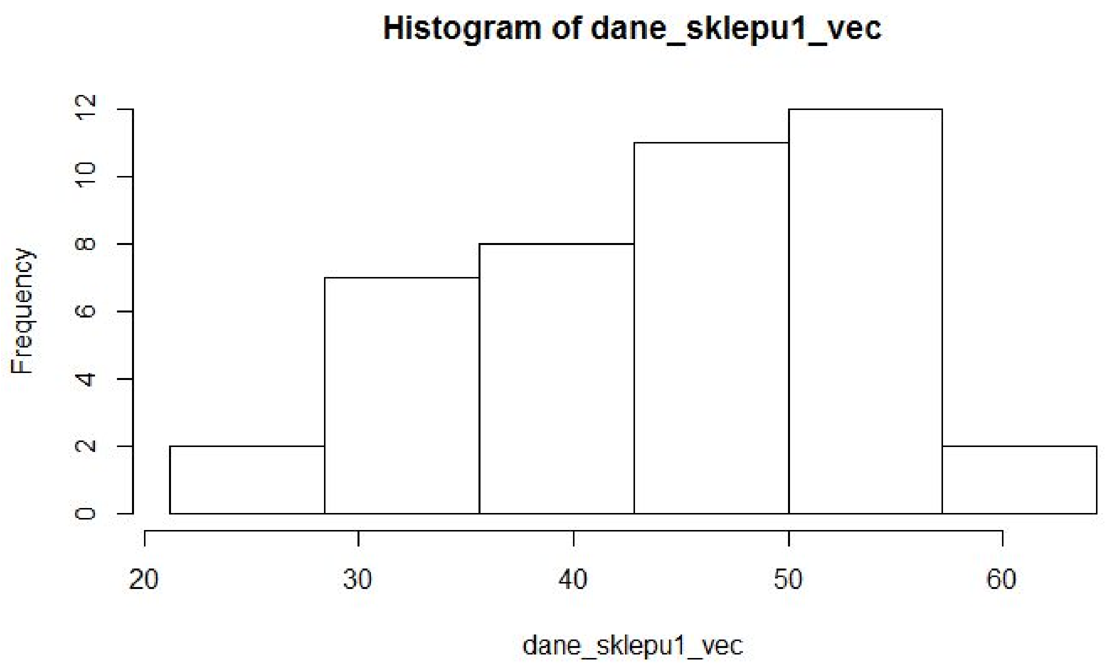
Przemyslaw Pawlas
Radosław Wojaczek
Zbigniew Kmonk
Rafał Grzelec
Dawid Kubów

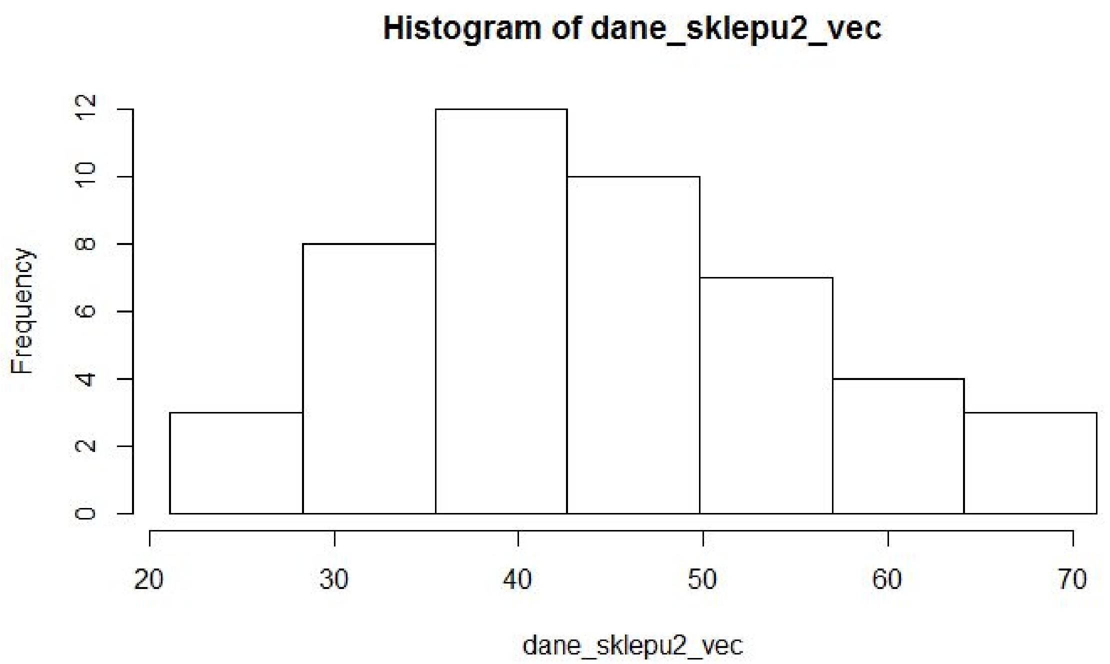
W poniższej tabeli zawarte są wyniki miar uzyskanych w trakcie realizacji zadań projektowych. Jest ona podzielona na szeregi oraz poszczególne sklepy.

	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 ćwiartkowe	7,365	7,96	7,569	7,794
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	454,997%	385,946%	475,49%	390,572%
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 dla poszczególnych sklepów utworzone przez środowisko R:





Zadanie 1

Wprowadziliśmy do programu dane, które są zawarte w treści zadania, dla każdego sklepu osobno. Korzystając z tych danych dokonaliśmy analizy statystycznej, wyznaczając miary przeciętne, zróżnicowania, asymetrii i koncentracji. Dane rozpatrzyliśmy w dwóch szeregach: szczegółowym oraz rozdzielczym. Dla każdego z nich przeprowadziliśmy obliczenia, a następnie zestawiliśmy uzyskane wyniki. Są one widoczne powyżej, wraz z histogramami.

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

Kolejnym zadaniem było przeprowadzenie testu Kołmogorowa – *Lillieforse'a*, który 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.

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

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

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

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

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

Ostatnie zadanie polegało na sformułowaniu i zweryfikowaniu hipotezy czy na poziomie istotnosci 0.05 mozna twierdzić, że wartość przeciętna miesięcznych wydatków na jedna osobę, na pieczywo i produkty zbożowe jest większa dla klientów pierwszego marketu.
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 - wariancje są różne
Wykorzystujemy do tego test Fishera o statystyce:
$$F = \frac{\text{wariancja nieobciążona pierwszego sklepu}}{\text{wariancja nieobciążona drugiego sklepu}}$$
oraz obszarze krytycznym:
$$(-\infty, f(1-\tfrac{\alpha}{2}, m-1,\ n-\ 1))$$
gdzie:
α - poziom istotności
m - liczebność danych pierwszego sklepu
n - liczebność danych drugiego sklepu
f - wartość odczytana funkcją qf()

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

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

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

gdzie:
$\overline{x}$ - średnia próby
s^2 - wariancja próby
i obszarze krytycznym:
$$(t(1 - \alpha, m + n - 2), \infty)$$
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:

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

Na podstawie wyznaczonych wartości oraz tego czy zawierają się w przedziałach można stwierdzić czy nie ma podstaw do odrzucenia hipotezy zerowej H_0.

Wyniki z konsoli po uruchomieniu programu:

Wynik testu: rowne odchylenia populacji, wiec: statystyka t = -0.011465, przedzial <1.662557, ∞). Wartosc nie nalezy do przedzialu - przyjmujemy hipoteze zerowa - wartości przeciętne sa rowne.

Wnioski:
Test na danym poziomie istotności wykazał, że wartości przeciętne dla obu sklepów są równe, więc nasze założenie 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.