2022

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

PROVOZNĚ EKONOMICKÁ FAKULTA

SEMESTRÁLNÍ PRÁCE ZE PŘEDMĚTU STATISTIKA II

OBSAH CUKRU V OVOCNÝCH KAPSIČKÁCH

OBSAH

1. ÚVOD 1
2. *VSTUPNÍ DATA*.....*3*
3. *METODIKA PRÁCE*.....*4*
4. PRŮZKUMOVÁ ANALÝZA DAT 5
5. REGERSNÍ A KORELAČNÍ ANALÝZA 8
6. REGRESNÍ DIAGNOSTIKA 10
7. SHRNUTÍ VÝSLEDKŮ 12
8. ZDROJE 12

1. ÚVOD

Ve své semestrální práci jsem se rozhodla porovnat, jestli existuje souvislost mezi obsahem cukru v ovocných kapsičkách a cenou. Cukr je obsažen v ovoci použitém k výrobě ovocných kapsiček, zajímá mě, zda obsah cukru, a tedy i ovoce, nějak ovlivňuje cenu, nebo vše záleží na konkrétní značce. Pokud existuje nějaká souvislost mezi obsahem cukru a cenou, tak bych mohla tuto analýzu použít při svých nákupech.

2. VSTUPNÍ DATA

Pro svou práci jsem použila data ze roku 2020 z českého kanálu A DOST! na webové stránce Stream.cz, kde byl proveden test ovocných kapsiček. Pro analýzu jsem vzala 14 druhů ovocných kapsiček a do tabulky jsem vypsala obsah cukru a cenu, za kterou se ovocné kapsičky prodávají v českých supermarketech.

	Název značky	**Obsah cukru (g/100g)**	**Cena za 100 g,Kč**
1	Ovocňák	14,1	20
2	RELAX	14,1	11,6
3	Kubík ovocná kapsička	14,5	12,9
4	HELLO ovocná přesnídávka	15,5	6,9
5	Sunárek	11,9	19,9
6	Hamé EASY FRUIT	12,4	22,6
7	Hamánek s JAHODAMI	10,9	16,6
8	dm babylove	10,5	16,9
9	peek-a-boo	10	30,9
10	hami Kouzelné Jablíčko	11,2	23,5
11	Bio lupilu	10,7	18,8
12	Nature's Promise	10,6	18,8
13	Nestlé NaturNes	11,1	24,3
14	Hipp Kinder	11,3	24,9

3. METODIKA PRÁCE

Pro tuto práci s daty využiji software SAS, ve kterém použiji k zjištění výsledků jednoduchý lineární model regresní a korelační analýzy. V práci budu pracovat s dvěma proměnnými. Cenou v Kč za 100 g ovocné kapsičky a obsahem cukru v g/100g.

4. PRŮZKUMOVÁ ANALÝZA DAT

Analýza – obsahu cukru

The SAS System

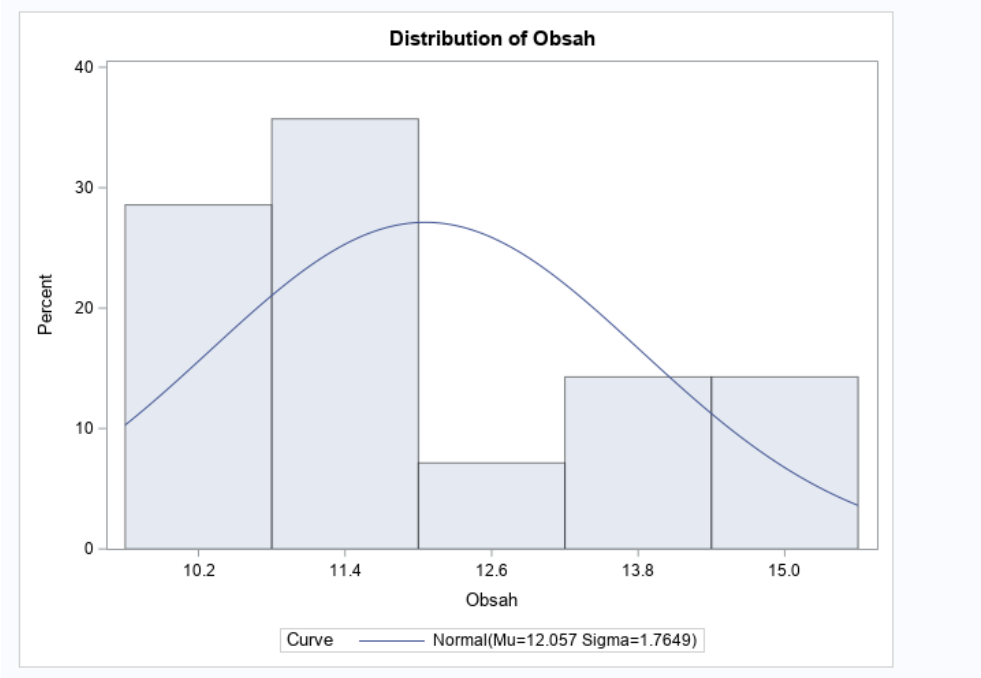
The UNIVARIATE Procedure
Variable: Obsah (Obsah)

Moments			
N	14	Sum Weights	14
Mean	12.0571429	Sum Observations	168.8
Std Deviation	1.76492069	Variance	3.11494505
Skewness	0.83104927	Kurtosis	-0.7363793
Uncorrected SS	2075.74	Corrected SS	40.4942857
Coeff Variation	14.6379678	Std Error Mean	0.4716949

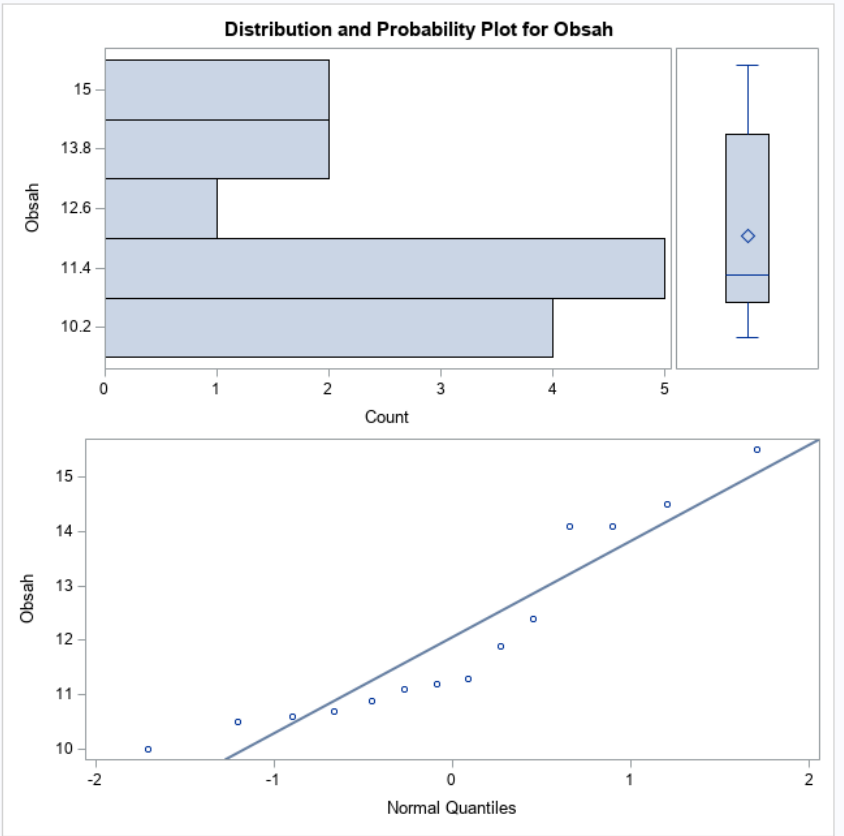
Průměrný obsah (MEAN) cukru v 14 kapsičkách je 12.05 g. Rozptyl (Variance) má hodnotu 3.11. Hodnota šikmosti (Skewness) je 0.83, značí mírnou pravostrannou asymetrii. Míra špičatosti je -0.7 a to ukazuje na plochost, což znamená, že mezi kapsičkami s nejvyšším a nejnižším obsahem cukru není velký rozdíl. Variační koeficient 14.63%, tzn., že se směrodatná odchylka podílí na aritmetickém průměru zhruba z patnácti procent.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.869911	Pr < W	0.0418
Kolmogorov-Smirnov	D	0.237465	Pr > D	0.0312
Cramer-von Mises	W-Sq	0.149416	Pr > W-Sq	0.0216
Anderson-Darling	A-Sq	0.826401	Pr > A-Sq	0.0243

K zjištění normality proměnné obsah využiji Shapiro-Wilkův test. Hodnota testu vyšla méně než 0,05. Proto se nulová hypotéza zamítá a proměnná obsah nemá normální rozdělení.



Když se podíváme na histogram, můžeme na něm vidět, že Gaussova křivka nekopíruje vrcholy. sloupců a potvrzuje nám, že proměnná obsah nemá normální rozdělení



Grafy potvrzují nenormálnost rozdělení u obsahu. Data jsou náhodná a jsou uspořádaná izolovaných shlucích.

Analýza – ceny ovocných kapsiček

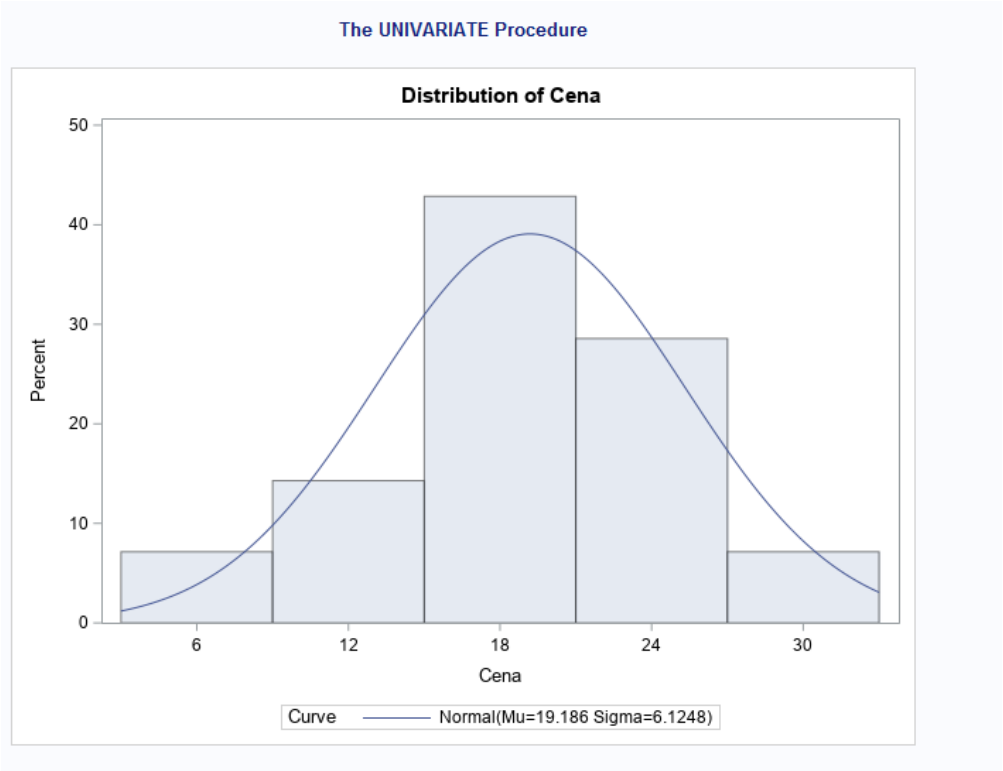
The UNIVARIATE Procedure
Variable: Cena (Cena)

Moments			
N	14	Sum Weights	14
Mean	19.1857143	Sum Observations	268.6
Std Deviation	6.12483684	Variance	37.5136264
Skewness	-0.2030037	Kurtosis	0.40197908
Uncorrected SS	5640.96	Corrected SS	487.677143
Coeff Variation	31.9239448	Std Error Mean	1.6369315

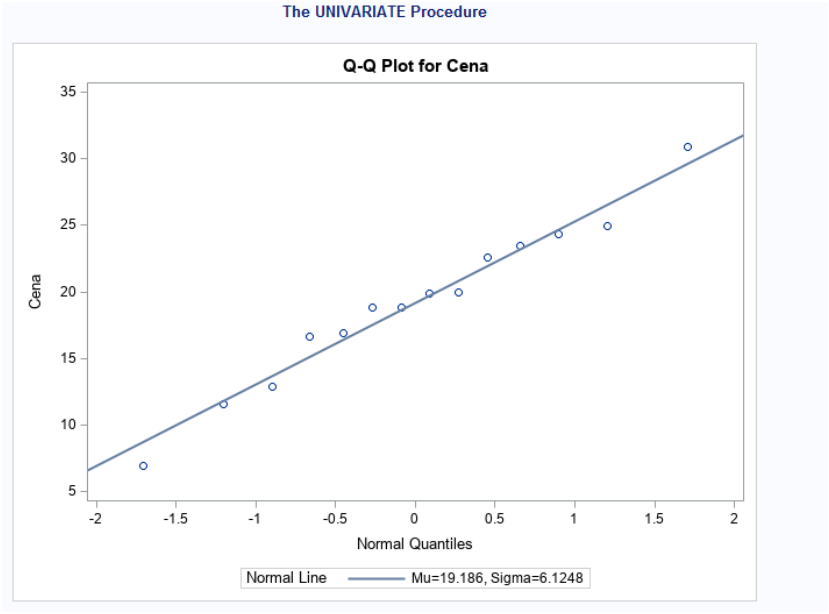
Průměrná cena 14 zkoumaných kapsiček je 19.18 Kč. Hodnota rozptylu je 37.51. Míra šikmosti má hodnotu -0.20, značí mírnou levostrannou asymetrii v rozložení četností. Hodnota špičatosti je 0.40, ukazuje rovnoměrnější rozdělení hodnot. Variační koeficient se rovná 31.9%, což znamená, že směrodatná odchylka podílí na aritmetickém průměru zhruba ze 32 procent.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.981787	Pr < W	0.9839
Kolmogorov-Smirnov	D	0.122165	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.031714	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.198452	Pr > A-Sq	>0.2500

Využiji Shapiro-Wilkův test normality u proměnné cena na rozdíl od proměnné obsah je větší než kritérium 0,05 (0.9839). Proto se nulová hypotéza přijímá a proměnná cena má normální rozdělení.



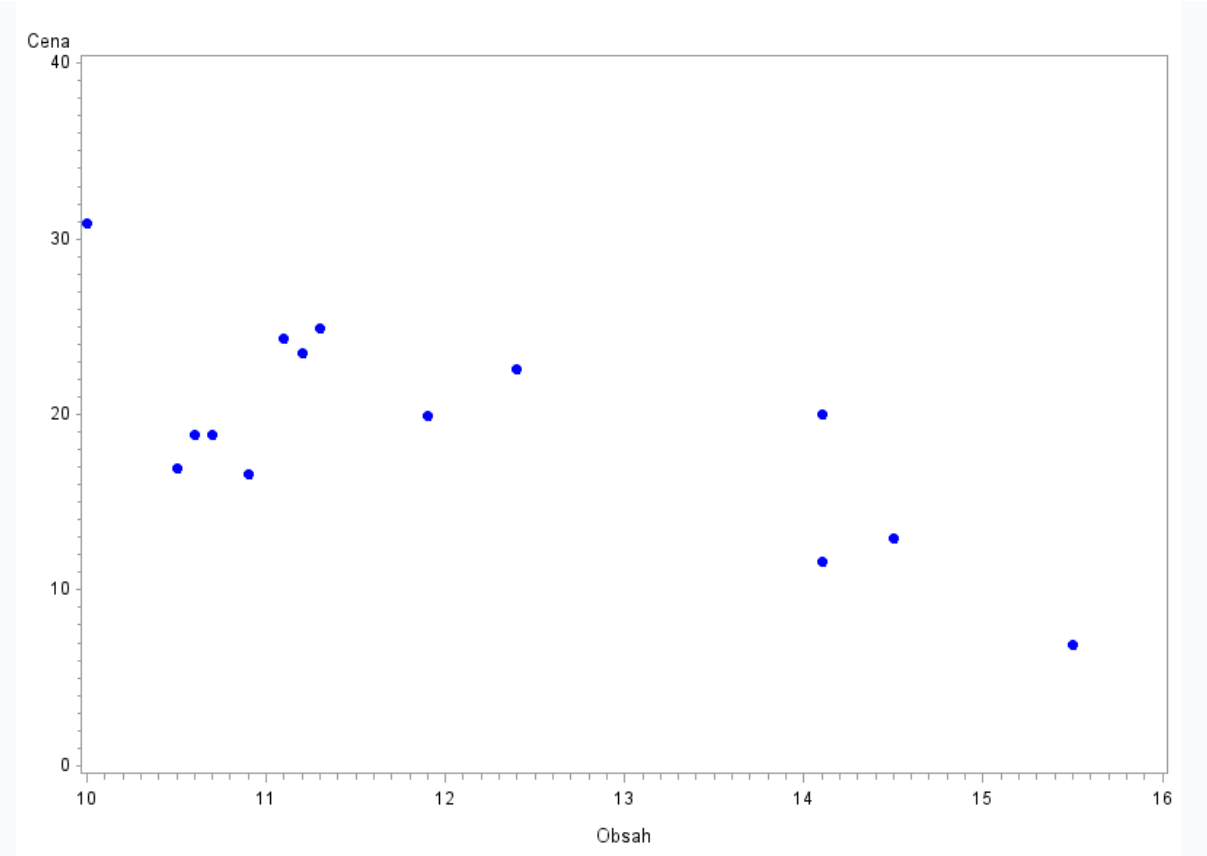
Gaussova křivka kopíruje vrcholy sloupců a potvrzuje nám, že proměnná obsah má normální rozdělení.



Vidíme, že body leží převážně na přímce s malými odchylkami podél každého z chvostů. Na základě tohoto grafu můžeme s jistotou předpokládat, že daný soubor dat má normální rozdělení.

5. REGRESNÍ A KORELAČNÍ ANALÝZA

Korelační pole



Korelační pole ukazuje, že mezi proměnnými cena a obsah existuje nepřímá (negativní korelace) závislost, čím vyšší je cena, tím méně sladké je ovocná kapsička, tj. nižší obsah cukru.

Analýza rozptylu

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	246.00387	246.00387	12.22	0.0044
Error	12	241.67327	20.13944		
Corrected Total	13	487.67714			

Z tabulky analýzy rozptylu můžeme vidět, že p-hodnota testu významnosti regresní funkce vyšla 0,0044, je tedy menší než 0,05. Tím pádem se nulová hypotéza zamítá, model je statisticky významný.

Těsnost závislostí

Root MSE	4.48770	R-Square	0.5044
Dependent Mean	19.18571	Adj R-Sq	0.4631
Coeff Var	23.39083		

Těsnost závislostí určuje koeficient determinace R-Square, který má hodnotu 0.5044. Korelační koeficient se vypočítá jako odmocnina 0.5044 a to se rovná 0.710211. Hodnota korelačního koeficientu je větší než 0.3, tudíž se jedná o střední silnou závislost mezi závislou proměnnou cena a nezávislou proměnnou obsah.

Odhady parametrů lineárního regresního modelu

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	48.90364	8.58715	5.69	<.0001
Obsah	Obsah	1	-2.46476	0.70522	-3.50	0.0044

Hodnota absolutního členu je 48.90364 a hodnota parametru obsah je -2.46476. Tvar regresní přímky je y = 48.90364 – 2.46476x, kde y – cena, x – obsah . Test významnosti pro absolutní člen je <.0001 a pro regresní koeficient je 0,0044. Jelikož je p-hodnota parametrů menší než α = 0,05, můžeme tento model považovat za významný.

Míra závislosti a těsnosti

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
Cena	14	19.18571	6.12484	19.35000	6.90000	30.90000	Cena
Obsah	14	12.05714	1.76492	11.25000	10.00000	15.50000	Obsah

Pearson Correlation Coefficients, N = 14
Prob > |r| under H0: Rho=0

	Cena	Obsah
Cena Cena	1.00000	-0.71024 0.0044
Obsah Obsah	-0.71024 0.0044	1.00000

Spearman Correlation Coefficients, N = 14
Prob > |r| under H0: Rho=0

	Cena	Obsah
Cena Cena	1.00000	-0.41189 0.1434
Obsah Obsah	-0.41189 0.1434	1.00000

Pearsonův (-0,71) i Spearmanův (-0,41) korelační koeficient značí negativní středně silnou lineární závislost. To potvrzuje předpoklad, nepřímé úměrnosti těchto proměnných – čím větší je obsah cukru, tím nižší bude cena ovocné kapsičky.

6. REGRESNÍ DIAGNOSTIKA

Analýza reziduí:

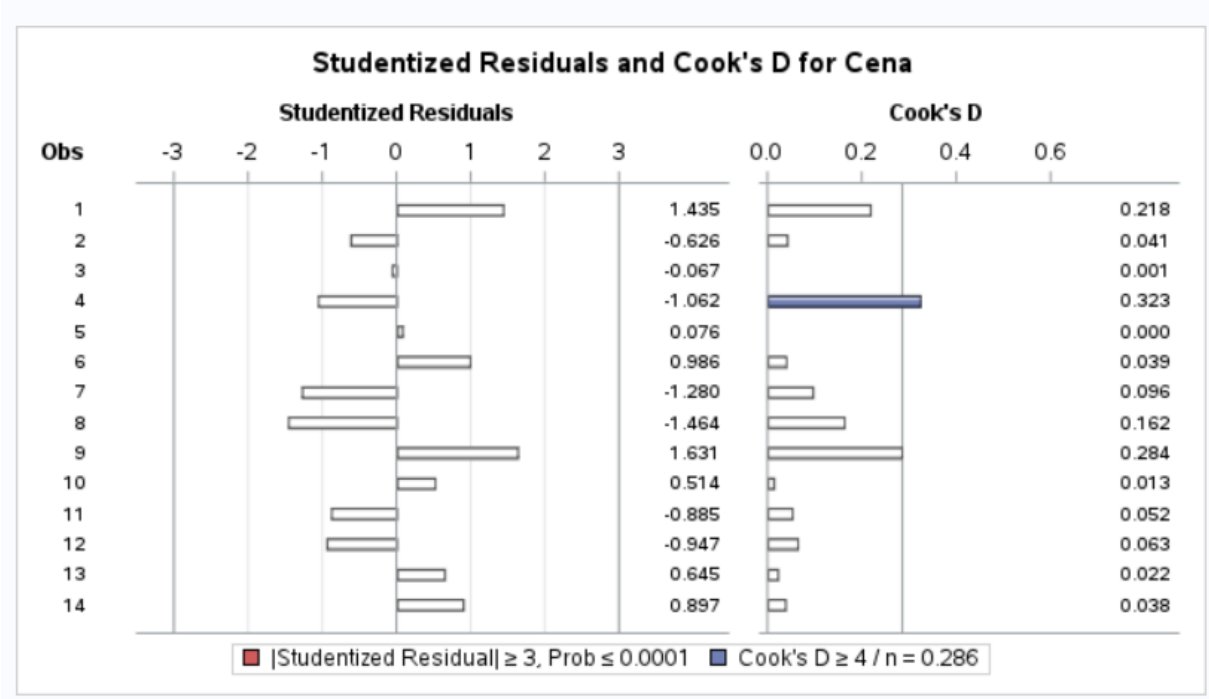
Whiteův test stálosti rozptylu reziduí:

The REG Procedure
Model: MODEL1
Dependent Variable: Cena Cena

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
2	2.17	0.3376

V tabulce máme výsledky Whiteova testu, který ověřuje splnění předpokladu homoskeadasticity reziduí neboli konstantnosti rozptylu reziduí. P-hodnota je větší než α = 0,05, tudíž se H0 přijímá. To v našem případě je, protože p-hodnota testu je 0,3376 a to znamená, že rezidua mají konstantní rozptyl.

Output Statistics													
												DFBETAS	
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	Obsah
1	20.0	14.1506	1.8746	5.8494	4.077	1.435	0.218	1.5090	0.1745	0.9895	0.6938	-0.4659	0.5332
2	11.6	14.1506	1.8746	-2.5506	4.077	-0.626	0.041	-0.6089	0.1745	1.3491	-0.2799	0.1880	-0.2151
3	12.9	13.1647	2.0992	-0.2647	3.966	-0.067	0.001	-0.0639	0.2188	1.5223	-0.0338	0.0248	-0.0278
4	6.9	10.6999	2.7081	-3.7999	3.579	-1.062	0.323	-1.0681	0.3641	1.5364	-0.8083	0.6676	-0.7247
5	19.9	19.5730	1.2045	0.3270	4.323	0.076	0.000	0.0724	0.0720	1.2812	0.0202	0.0046	-0.0019
6	22.6	18.3407	1.2235	4.2593	4.318	0.986	0.039	0.9853	0.0743	1.0856	0.2792	-0.0164	0.0552
7	16.6	22.0378	1.4507	-5.4378	4.247	-1.280	0.096	-1.3194	0.1045	0.9906	-0.4507	-0.3031	0.2535
8	16.9	23.0237	1.6262	-6.1237	4.183	-1.464	0.162	-1.5466	0.1313	0.9243	-0.6013	-0.4640	0.4061
9	30.9	24.2561	1.8823	6.6439	4.074	1.631	0.284	1.7699	0.1759	0.8749	0.8178	0.6969	-0.6303
10	23.5	21.2984	1.3431	2.2016	4.282	0.514	0.013	0.4978	0.0896	1.2502	0.1561	0.0891	-0.0703
11	18.8	22.5307	1.5345	-3.7307	4.217	-0.885	0.052	-0.8760	0.1169	1.1776	-0.3187	-0.2317	0.1988
12	18.8	22.7772	1.5794	-3.9772	4.201	-0.947	0.063	-0.9424	0.1239	1.1630	-0.3543	-0.2659	0.2305
13	24.3	21.5448	1.3763	2.7552	4.271	0.645	0.022	0.6285	0.0941	1.2241	0.2025	0.1230	-0.0993
14	24.9	21.0519	1.3129	3.8481	4.291	0.897	0.038	0.8888	0.0856	1.1329	0.2719	0.1442	-0.1106



Vybočující hodnoty

Sloupec Hat Diag H – diagonální prvky projekční matice nám říká, jestli se v množině vysvětlující proměnné objevuje vybočující hodnota. Hranice pro vybočující hodnotu se vypočítá 2*(2/14) a to se rovná 0,286. Tuto hodnotu překračují pozorování 4, jedná se tedy o vybočující hodnoty.

Odlehlé hodnoty

Sloupec Student Residual nám říká, jestli se v množině vysvětlované proměnné objevuje odlehlá hodnota. Absolutní hodnota odlehlé hodnoty musí být větší než 2. Tato hodnota není překročena pozorováním.

Vlivné pozorování

Vlivné hodnoty můžeme určit pomocí sloupce Cook's D nebo sloupce DFFITS. Kritérium pro posouzení celkové míry vlivnosti pozorování u sloupce Cook's D je 4/14 = 0,286. Hranici překračuje stejně jako u studentizovaných reziduí pozorování 4. Jedná se tedy o vlivnou hodnotu. Hranice u sloupce DFFITS pro určení míry vlivnosti vyrovnané hodnoty se vypočítá jako absolutní hodnota 2* odmocnina (2/14) = 0,756. Tuto hranici překračuje pozorování 9 a jedná se o vlivnou hodnotu.

7. SHRNUTÍ VÝSLEDKŮ

Hlavním cílem mé práce bylo zjistit, jestli je cena ovocných kapsiček závislá na obsahu cukru. Průzkumová analýza potvrdila normalitu pouze u proměnné cena. Pomocí korelačního pole jsem zjistila, že mezi proměnnými je nepřímá lineární závislost. Intenzita závislosti mezi proměnnou cena a proměnnou obsah je 50% a střední silná závislost s hodnotou 0.710211. Rezidua mají konstantní rozptyl. Podle regresní analýzy a odhadů parametrů lineárního regresního modelu jsou výsledky statisticky významné, a proto byla závislost mezi cenou a obsahem statisticky prokázaná. Díky této práci jsem si uvědomila, že čím nižší cena, tím více cukru ve výrobku, který s největší pravděpodobností působí jako konzervant pro delší skladování, a tím se výrobek stává méně zdravým a méně kvalitním.

8. ZDROJE

- statistický program SAS
- https://www.stream.cz/adost/kompletni-vysledky-nezavislych-testu-ovocnych-kapsicek-64102772