VILNIAUS UNIVERSITETO

KAUNO FAKULTETAS

SOCIALINIŲ MOKSLŲ IR TAIKOMOSIOS INFORMATIKOS INSTITUTAS

Informacinių sistemų ir kibernetinės saugos studijų programa

IRMANTAS VASILJEVAS

STATISTIKOS STUDIJŲ DALYKAS

SAVARANKIŠKAS DARBAS No. 2

KAUNAS 2019

TURINYS

PAVEIKSLŲ SĄRAŠAS

LENTELIŲ SĄRAŠAS

ĮŽANGA

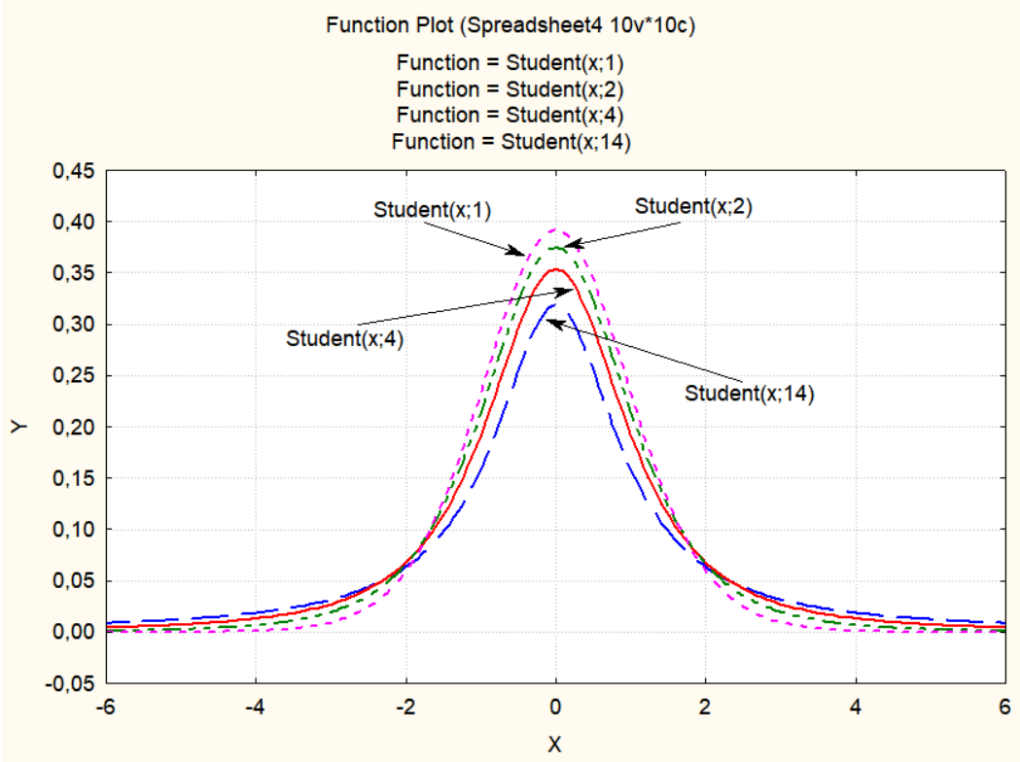
Šio darbo tikslas yra pademonstruoti Statisitikos studijų dalyko praktinių ir teorinių užsiėmimų metu įgytas žinias bei įgūdžius.

1. PIRMOJI UŽDUOTIS

Nubrėžti bent trijų pasiskirstymų grafikus (bent vienas dviejų parametrų, bent du tolydiniai). Vieno pasiskirstymo vaizdavimui naudoti bent po keturias pasiskirstymo kreives, gaunamas skirtingoms parametrų reikšmėms (bent vieno parametro reikšmė turi sutapti su jūsų Nr.). Paaiškinti kitimo tendencijas kintant parametrams.

1.1 Student(x;df) funkcijos grafikai

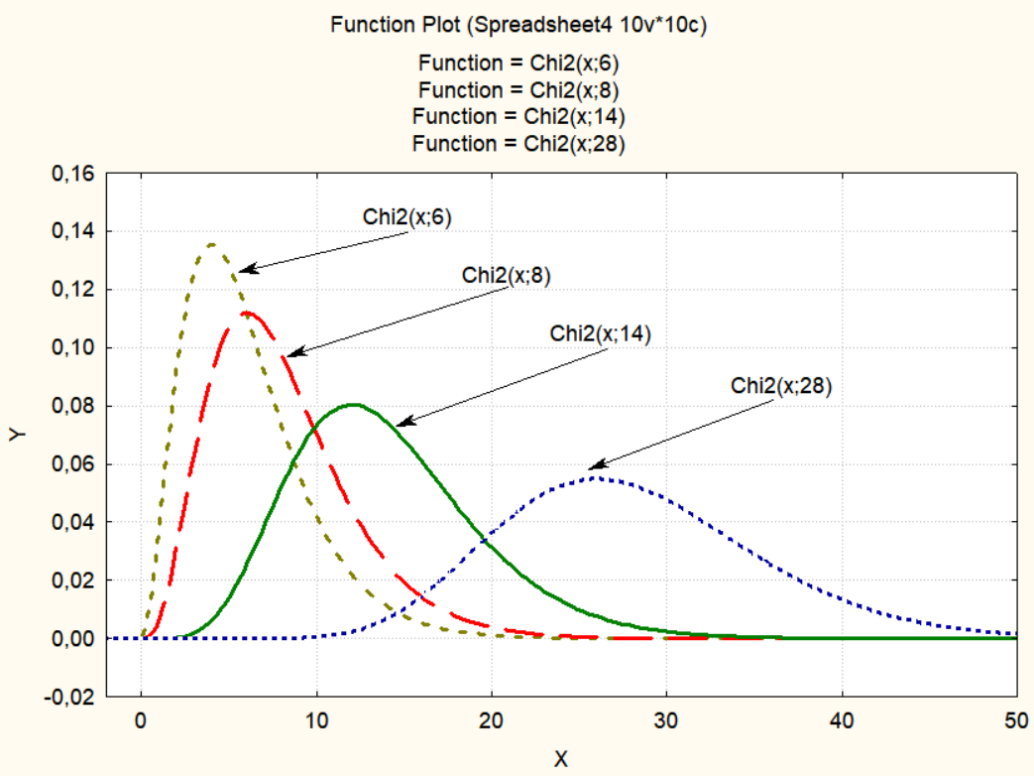
Student() funkcijos grafikas redaguojant df parametrą kinta grafiko viršūnės aukštis Y ašies atžvilgiu. Kai parametro reikšmė didėja, viršūnės aukštis mažėja ir atvirkščiai, kai parametro reikšmė mažėja, viršūnės aukštis didėja. Be to, kai parametro reikšmė didėja, grafiko šakos didesniu kampu artėja prie begalybės, kai mažėja - mažesniu.



pav 1 Student() funckijų grafikai

1.2 X^2 - Chi2(x;nu) - funkcijos grafikai

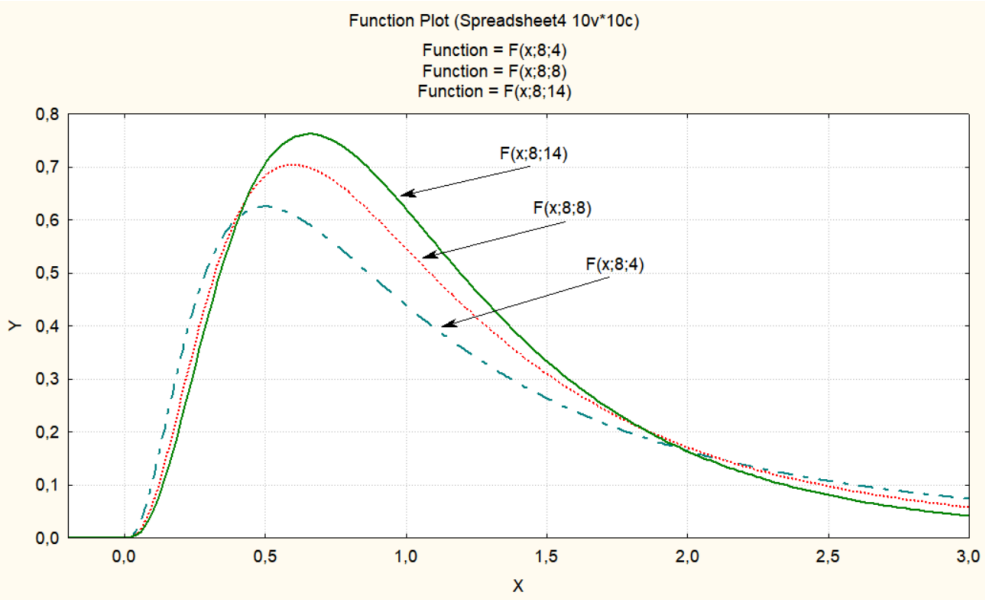
Chi2() funkcijos grafikas primena pusę sinusoidės, išsitęsiančios iki plius ir minus begalybės, teigiamoje Y ašyje. Didėjant parametro nu reikšmei grafiko viršūnė žemėja, bet kartu platėja grafiko šakos.



pav 2 Chi2() funkcijų grafikai

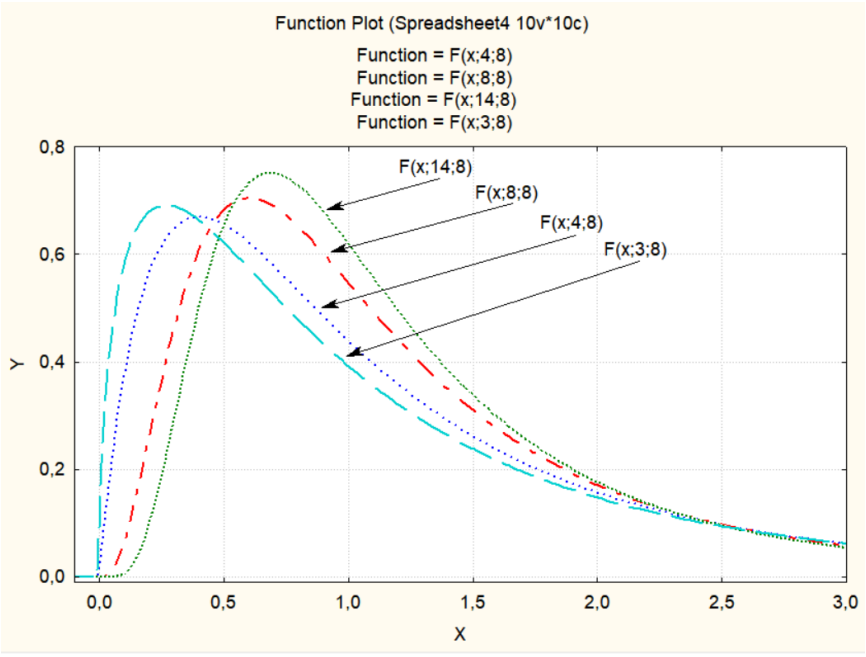
1.3 F(x;nu;omega) funkcijos grafikai

F() funkcija turi du parametrus. Kai mažėja antrasis parametras omega, tada vyksta grafiko transformacija link neigiamos Y ir X ašies pusių, tai reiškia, jog grafiko viršūnė žemėja Y ašies atžvilgiu bei grafiko šakos artėja link Y ašies. Be to, didėjant parametro reiškmei didėja grafiko šakų leidimosi kampas.



pav 3 F() funkcijų grafikai I

Didėjant parametrui nu vyksta transformacija įstryžai nuo Y ir X ašių į teigiamą pusę. Kai parametro reikšmė mažėja iki tam tikro lygio, tai atitinkamai vyksta grafiko transformacija įstrįžai link Y ir X ašių neigiamųjų pusių, tai yra, grafiko viršūnės X_0 ir Y_0 įgauna mažesnes reikšmes. Parametro reikšmei mažėjant bei pasiekus tam tikrą kritinę reikšmę, grafiko viršūnė pradeda kilti Y ašies atžvilgiu, bet vis tiek mažėja X ašies atžvilgiu ne arčiau kaip iki $X_0 = 0$.



pav 4 F() funkcijų grafikai II

2. ANTROJI UŽDUOTIS

Pagal pirmo savarankiško darbo (pirmos užduoties) statistinius duomenis, surasti vidurkio intervalinį įvertinimą dviem skirtingais būdais (p=0,95+nr/500). Patikrinti hipotezę apie tų duomenų normališkumą. Parinkti geriausią intervalų skaičių.

2.1 Imties vidurkio intervalinis įvertinimas generuojant automatiškai

Imties vidurkio intervalinis įvertinimas gali būti sugeneruojamas su duomenų sklaidos bei padėties charakteristikomis STATISTICA programoje. Renkamės meniu juostos punktą *Statistics*, po to *Basic Statistics/Tables*. Iššokusiame dialogo lauke tarp įvairių duomenų vaizdavimo galimybių reikia pasirinkti *Descriptive Statistics* bei paspausti mygtuką *OK*. Po to, sklityje *Advanced* pakeičiame *Conf. limit for means* reikšmę į 95, 8 %, kadangi to reikalauja užduoties sąlyga. Paspaudus mygtuką *Summary* naujame lange sugeneruojami sklaidos ir padėties charakteristikos, tarp kurių jau yra imties vidurkio intervalinis įvertinimas. Mūsų atveju jis apima intervalą - (159180,76; 166003,50).

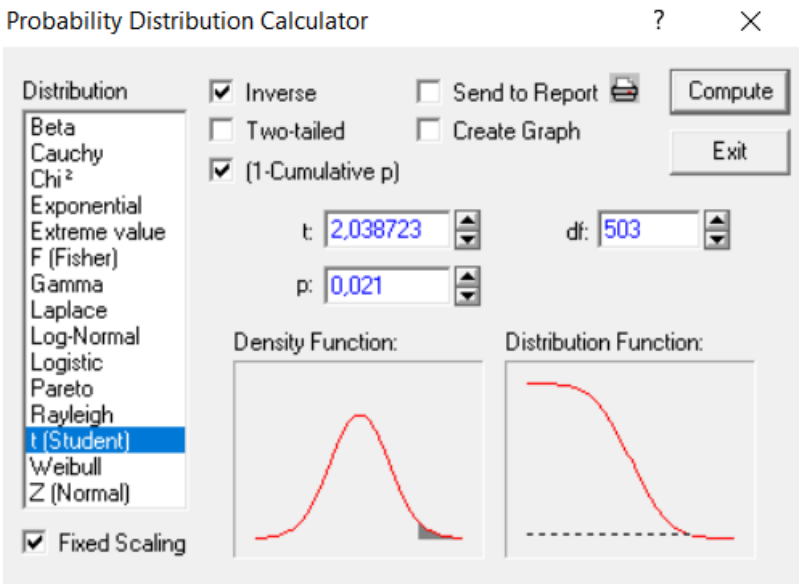
Descriptive Statistics (Spreadsheet4 in Statistika-savr-2)

Variable	Valid N	Mean	Confidence -95,800%	Confidence +95,800%	Std.Dev.
Var1	504	162592,13112731479	159180,75986735168	166003,5023872779	37565,222893372251

pav 5 Imties duomenų sklaidos ir padėties charakteristikos

2.2 Imties vidurkio intervalinis įvertinimas skaičiuojant pagal formulę

Imties intervalinis vidurkio įvertinimas skaičiuojamas pagal formulę - $X \pm t * \frac{S}{\sqrt{n}}$, kur X yra imties vidurkis, t - kvartilis, S - standartinis nuokrypis, o n - imties dydis. X, S, ir n dydžius gauname iš sugeneruotų duomenų sklaidos bei padėties charakteristikų. Kintamojo t reikšmę gauname pasinaudoję *Propability Calculator t(Student)* funkcijos skaičiuokliu, kuriame $df = n - 1$, o $p = \frac{1-C}{2}$, C yra *Conf. limit for means* dešimtainė išraiška. Mūsų atveju $p = \frac{1-0.958}{2} = 0.021$.



pav 6 Tikimybių skaičiuotuvas

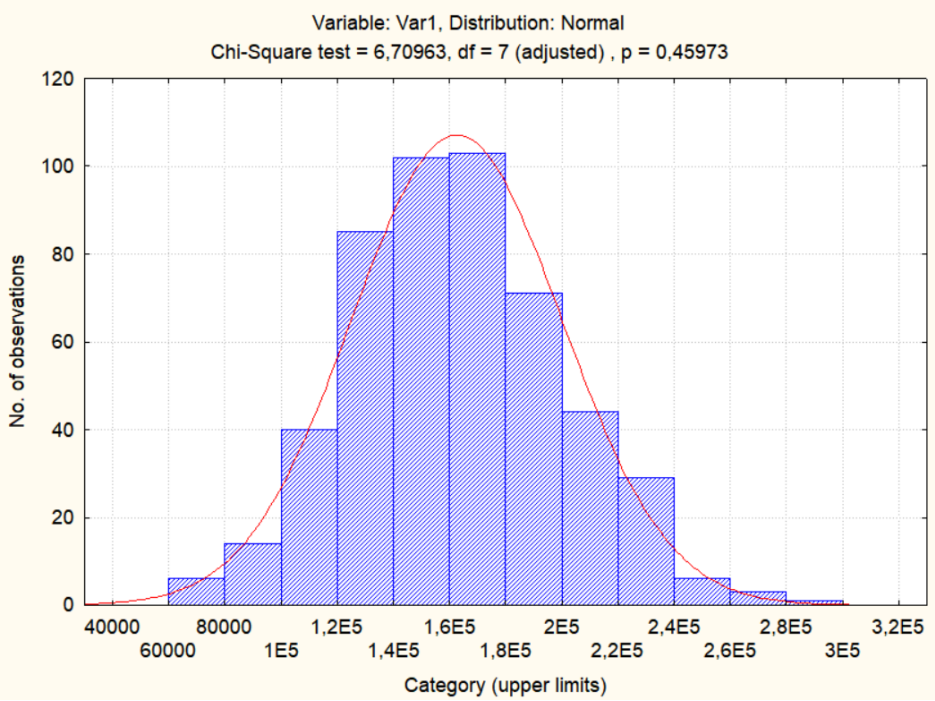
Suskaičiavus pagal formulę imties vidurkio intervalinis įvertinimas atitinka intevalą (159181,67; 166004,59).

Lentelė 1 Imties vidurkio intervalinis Įvertinimas pagal formulę

	Imtis	
Standartinis nuokrypis (S)	37566, 2228933723	
Vidurkis (X)	162593,131127315	
Skirstinio dydis (N)	504	
Kvartilis (t)	2,038723	
	Intervalas	Formulė
Min	159181,6695	X-t*S/SQRT(N)
Max	166004,5927	X+t*S/SQRT(N)

2.3 Imties duomenų normališkumo tikrinimas

Norėdami patikrinti imties duomenų normališkumą nusibrėšime grafiką, iš kurio matysis, kaip išsidėstę duomenys. Meniu juostoje renkamės *Statistics*, susirandame punktą *Distribution Fitting*, iššokusiame lange pasirenkame *Distribution -> Normal*. Kitame iššokusiame lange prisiskiriame kintamąjį, kuriame yra mūsų duomenys bei spaudžiame mygtuką *Plot of observed and expected distribution*. Gauname sekantį grafiką, kuriame iš kreivės bei histogramos matosi, jog duomenys yra normališkai išsidėstę.



pav 7 Duomenų normališkumą parodanti diagrama

3. TREČIOJI UŽDUOTIS

Iš statistinių žinynų ar periodinių leidinių (pateikti šaltinio bibliografinį aprašymą) surasti tris skirtingus duomenų blokus ir pritaikyti :

t-kriterijų priklausomiems

t-kriterijų nepriklausomiems duomenims

x^2 kriterijų (n+m>4)

T kriterijus priklausomoms imtims

Duomenų analizės situacija - turime dvyliką žmonių. Yra matuojamas jų kraujo spaudimas du kartus dvejose skirtingose pozicijose - stovint ir gulint.

H_0 - kraujo spaudimo vidurkiai matuojant esant skirtingoms kūno padėtims yra vienodi

	Kraujo spaudimas	
	1 Stovint	2 Gulint
1	132	136
2	146	145
3	135	140
4	141	147
5	139	142
6	162	160
7	128	137
8	137	136
9	145	149
10	151	158
11	131	120
12	143	150

pav 8 Kraujo spaudimo rodmenys

Kadangi klaidos tikimybė p = 0,143790, o tai yra daugiau už 0,05, dėl to turime atmesti hipotezę, kad kraujo spaudimo vidurkiai matuojant esant skirtingoms kūno padėtims yra vienodi, tad priimame hipotezę H_1, pagal kurią vidurkiai skiriasi.

Variable	T-test for Dependent Samples (Spreadsheet53 in Statistika-savr-2.stw) Marked differences are significant at p < ,05000							
	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
Stovint	140,8333	9,49482						
Gulint	143,3333	10,83205	12	-2,50000	5,502066	-1,57400	11	0,143790

pav 9 T-kriterijus priklausomoms imtis

Lentelių generavimo scenarijus -

Statistics -> Basic Statistics/Tables -> t-tests, dependent samples -> OK -> Variables -> Summary

T kriterijus nepriklausomoms imtims

Duomenų imtis sudaro dviejų skirtingų vilkikų markių mažmeninės kainos. Taikydami T kriterijų sieksime nustatyti, ar skirtingos vilkikų markės lemia vilkikų kainas.

H_0 - skirtingų markių vilkikų mažmeninis kainų vidurkis yra vienodas

	Skirtingų markių vilkikų kainos	
	1 Pikapas 1	2 Pikapas 2
1	17400	17500
2	23300	23700
3	29200	20800
4	19200	22500
5	17600	24300
6	19200	26700
7	23600	24500
8	19500	17800
9	22200	29400
10	24000	29700
11	26400	20100
12	23700	21100
13	29400	22100
14	23700	24200
15	26700	27400
16	24000	28100

pav 10 Vilkikų kainos

Kadangi klaidos tikimybė p = 0,616697, o tai yra daugiau už 0,05, dėl to turime atmesti hipotezę, kad skirtingų markių vilkikų mažmeninis kainų vidurkiai yra vienodi. Tai reiškia, kad skirtingų markių vilkikų mažmeniniai kainų vidurkiai skiriasi.

	T-test for Independent Samples (Spreadsheet11 in Statistika-savr-2.stw) Note: Variables were treated as independent samples										
Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
Pikapas 1 vs. Pikapas 2	23068,75	23743,75	-0,505796	30	0,616697	16	16	3744,279	3804,728	1,032550	0,951357

pav 11 T-kriterijus nepriklausomoms imtis

Lentelių generavimo scenarijus -

Statististics -> Basic Statistics/Tables -> t-tests, independent by variables -> OK -> Variables -> Summary

Chi2 kriterijus

Duomenų pasirinkimo situacija - 400 mokinių buvo suskirstyti į dvi grupes pagal asmenybės tipus į intarvertus ir ekstravertus. Be to, kiekvienas iš mokinių pasirinko savo mėgstamiausią spalvą iš keturių duotųjų - raudonos, geltonos, žalios, mėlynos.

Taikydami Chi2 kriterijų sieksime nustatyti, ar asmenybės tipas lemia spalvų pasirinkimą.

H_0 - intravertai ir ekstravertai renkasi vienodas spalvas

	1 Raudona	2 Geltona	3 Žalia	4 Mėlyna
Intravertai	20	6	30	44
Ekstravertai	180	34	50	36

pav 12 Spalvų pasirinkimas pagal asmenybės psichotipą

Kadangi klaidos tikimybė p yra mažesnė už 0,05, tai hipotezės negalime atmesti, hipotezė yra teisinga.

Test	Table: TIPAS(2) * SPALVA(4) (Spreadsheet36) Model: 1,2		
	Chi-sqr	df	p
Max Likelihood Chi-square	56,04453	3	0,00000000000411005
Pearson Chi-square	56,81452	3	0,00000000000281519

pav 13 Chi2 spalvų pasirinkimo hipotezei

Lentelių generavimo scenarijus -

Statistics -> Advanced Linear/Nonlinear Models -> Log-Linear Analysis of Frequency Tables -> Input file: Frequencies without coding variables -> Variables -> Select All -> Specify table (2 : TIPAS; 4 : Spalva) -> OK -> Specify model to be tested (1 2) -> OK -> Summary

4. KETVIRTOJI UŽDUOTIS

Sugalvoti situaciją dispersinės analizės pritaikymui. Išanalizuoti ir išspręsti vieno faktoriaus situaciją.

Duomenų pasirinkimo situacija

Šioje užduotyje yra nagrinėjama, ar skiriasi skrydžių trukmė skirtingos avialinijose renkantis maršrutą Vilnius-Oslas. Viename duomenų stulpeyje pateiktas skrydžių laikas minutėmis, kitame viena iš trijų avialinijų.

H_0 - avialinijų skrydžių trukmė tais pačiais maršrutais yra vienoda

	1 Laikas	2 Avialinija
1	96	b
2	97	b
3	102	c
4	102	a
5	103	b
6	103	a
7	104	a
8	106	b
9	107	c
10	109	c
11	109	c
12	112	b
13	113	a
14	116	a
15	119	a
16	119	c
17	120	a
18	120	a
19	121	b
20	126	c
21	129	b
22	132	b
23	136	c
24	139	c

pav 14 Skrydžių laikai

Kadangi klaidos tikimybė p = 0,492822, o tai yra daugiau už 0,05, dėl to turime atmesti hipotezę, kad skirtingų avialinijų skrydžių trukmė tais pačiais maršrutais yra vienoda, tai reiškia, kad skrydžių trukmė skiriasi.

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Laikas	220,1871	2	110,0936	3158,442	21	150,4020	0,731995	0,492822

Analysis of Variance (Spreadsheet5 in Statistika-savr-2.stw)
Marked effects are significant at p < ,05000

pav 15 Skrydžių trukmės analizė

Lentelėje žemiau matome lentelėje, kurioje pavaizduoti skrydžių laiko vidurkiai trejose avialinijose.

Breakdown Table of Descriptive Statistics (Spreadsheet5 in Statistika-savr-2.stw)
N=24 (No missing data in dep. var. list)

Avialinija	Laikas Means	Laikas N	Laikas Std.Dev.
a	112,0797	8	7,64003
b	111,8672	8	14,12257
c	118,3961	8	13,90644
All Grps	114,1143	24	12,12010

pav 16 Skrydžių vidurkiai trijose avialinijose tuo pačiu maršrutu

Lentelių generavimo scenarijus -

Statistics -> Breakdown & one-way ANOVA -> OK -> Variables (Laikas, Avialinija) -> OK -> Codes for grouping variables -> All -> OK -> Lists of tables -> Output Tables -> Analysis of Variance -> OK -> Summary

IŠVADOS

X^2 kriterijaus taikymas, T-kriterijaus taikymas priklausomoms ir nepriklausomoms duomenų imtims, regresijos taikymas gali padėti atrasti koreliacijas tarp duomenų bei paneigti arba patvirtinti hipotezes apie duomenis. Šios įžvalgos gali būt naudojamos priimant verslo sprendimus.

ŠALTINIAI

1. Virgilijus Sakalauskas. *Statistikos paskaitų medžiaga*. Žiūrėta 2019-11-21 per internetą: VU informacijos šaltiniai.

2. Cengage Learning. *Data Sets, Pickup Trucks*. Žiūrėta 2019-12-01 per internetą: <http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/tvis/frames/frame.html>.

3. Learntech UNW Bristol. *Paired t-tests.* Žiūrėta 2019-12-01 per internetą: <http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1439>.

4. Learntech UNW Bristol. *Chi-squared test for nominal (categorical) data.* Žiūrėta 2019-12-01 per internetą: <http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1435>.