Dávid Nikolett
DANRABT.SZE

Statisztikai szoftverek esszé

Dávid Nikolett

Szeged
2011

1. Helyzetfelmérés

Adott egy kölcsön.txt nevű adatfájl, amely információkkal rendelkezik az ügyfelek életkoráról, családi állapotáról, munkaviszonyáról, a kölcsön típusáról, a kölcsönt felvevő éves jövedelméről, a fedezetről, a kölcsön és a fedezet hányadosáról (kperf), az adóstípusról, és a státuszról.

2. Az elemzés célja

Elemzésem céljai a következők:

1. Megvizsgálni, hogy életkorban van-e különbség a családi állapot, valamint az ügyfél statusa vonatkozásában.
2. Annak vizsgálata, hogy a munkaviszony és a kölcsönt felvevők éves jövedelme korrelál-e az életkorral.
3. Kíváncsiak vagyunk arra is, hogy a kölcsönt felvevők éves jövedelme az egyes családi állapotok szerint különbözik-e.

Munkámat az általam választott SPSS 17.0 programmal fogom elvégezni.

3. Az adatok megtekintése

Az adatelemzés első lépése az adatok megtekintése. Tehát az alábbi adatok állnak a rendelkezésünkre:

- csalall: 1. egyedülálló, 2. elvált vagy özvegy, 3. házas
- ekor: életkor
- munkvisz: hány éve dolgozik a jelenlegi munkahelyén
- típus: 1. beruházási kölcsön, 2. ingatlanvásárlási kölcsön, 3. vásárlási kölcsön
- jovedelem: a kölcsönt felvevő éves jövedelme
- fedezet: 1. vegyes, 2. autó, 3. ingatlan, 4. részvény, 5. fedezetlen
- kperf: kölcsön és fedezet hányadosa
- adostip: a kölcsönt felvevő ezt megelőzően volt-e ügyfele a banknak; 1. nem, 2. igen
- status: 1. problémamentes törlesztés, 2. törlesztési problémák előfordulnak, 3. fizetésképtelen

4. Deskriptív statisztika

Most még csak változónként nézzük meg az adatokat. Egy változó értékeiről azt érdemes tudni, hogy mely érték hányszor következett be, vagyis, hogy, hogy oszlanak el az adatok egy mintán belül. Az a cél, hogy kialakuljon bennünk egy kép az adatokról. Az adatok rendszerezésének legegyszerűbb módja a gyakoriság táblázat, amelyet az Analyze → Destriptive Statistics → Frequencies menüponttal érhetünk el. Tekintsük először a nominális változókat, a családi állapotot és az ügyfél statusát. Ekkor a gyakoriság táblázatok így néznek ki:

csalall

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	egyedülálló	16	22,9	22,9	22,9
	elvált vagy özvegy	17	24,3	24,3	47,1
	házas	37	52,9	52,9	100,0
	Total	70	100,0	100,0	

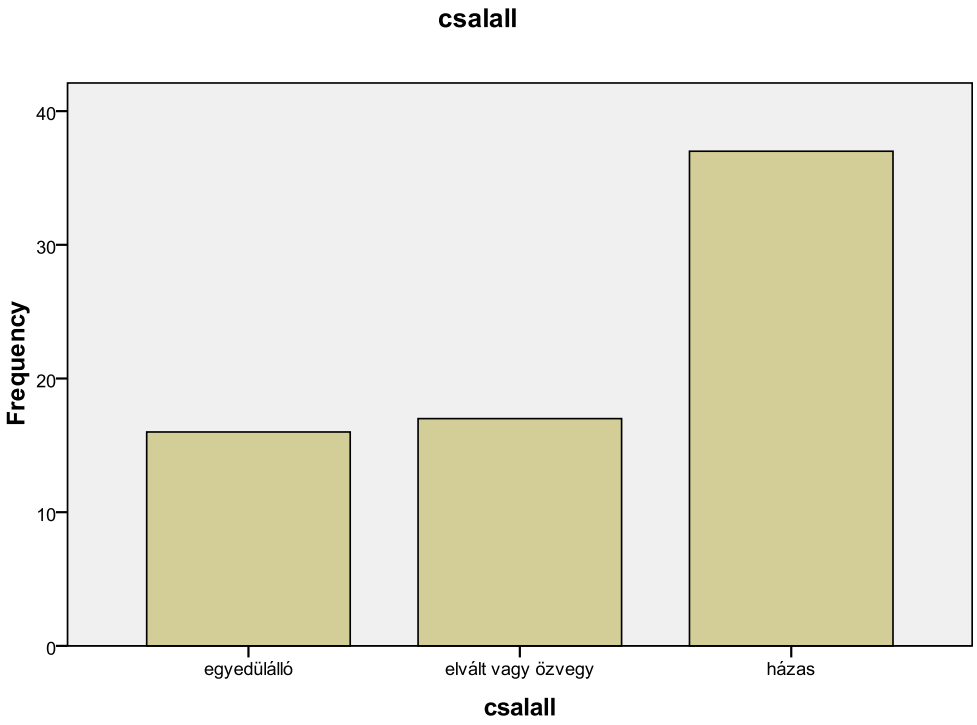
Látható például, hogy 37 házas ember van a mintában és ez a 70 ember 52,9 %-a. Még azt érdemes megemlíteni, hogy itt a sorrend nem számít, tehát nincs értelme kijelenteni, hogy a gyakoriságok monoton növekednek. Ugyanis bármilyen sorrendben is szerepelhetnének.

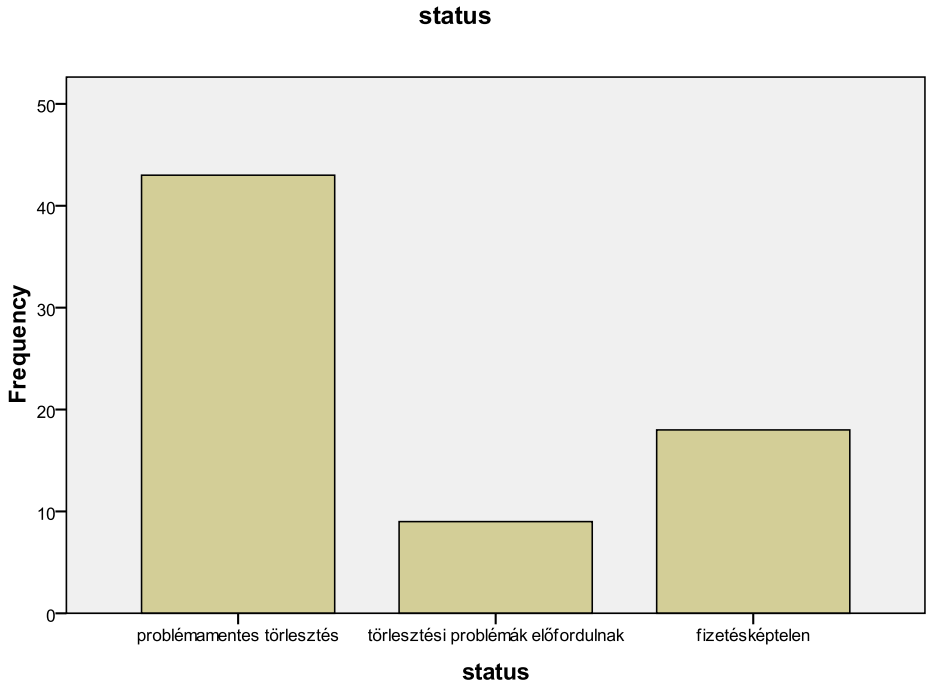
Status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	43	61,4	61,4	61,4
	2	9	12,9	12,9	74,3
	3	18	25,7	25,7	100,0
	Total	70	100,0	100,0	

A státusznál pedig megállapíthatjuk, hogy 43 személy problémamentesen törleszt, ami a 70 fő 61%-a. A sorrend ezen esetben sem számít.

Sokkal könnyebb benyomást szerezni az adatokról, ha grafikusan jelenítjük meg őket. Ezeket az ábrákat oszlopdiagramnak szokás nevezni. A gyakoriságok a következő módon szemléltethetők:


csalall
Frequency
40
30
20
10
0
egyedülálló
elvált vagy özvegy
házas
csalall


status
Frequency
50
40
30
20
10
0
problémamentes törlesztés
törlesztési problémák előfordulnak
fizetésképtelen
status

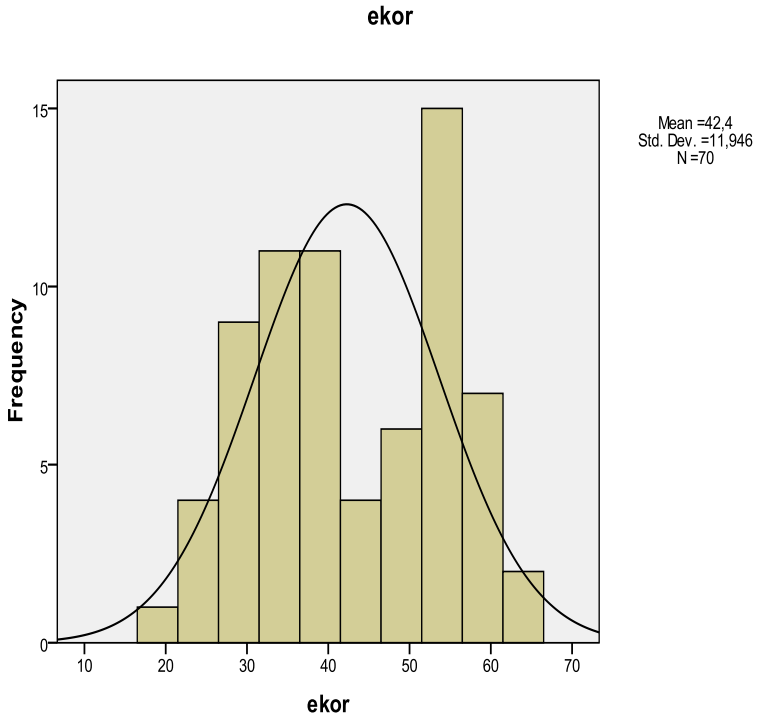
Most tekintsük folytonos változóinkat: az életkort, jövedelmet, és a munkaviszonyt. Először itt is deskriptív statisztikát készítünk az adatokról.

Statistics

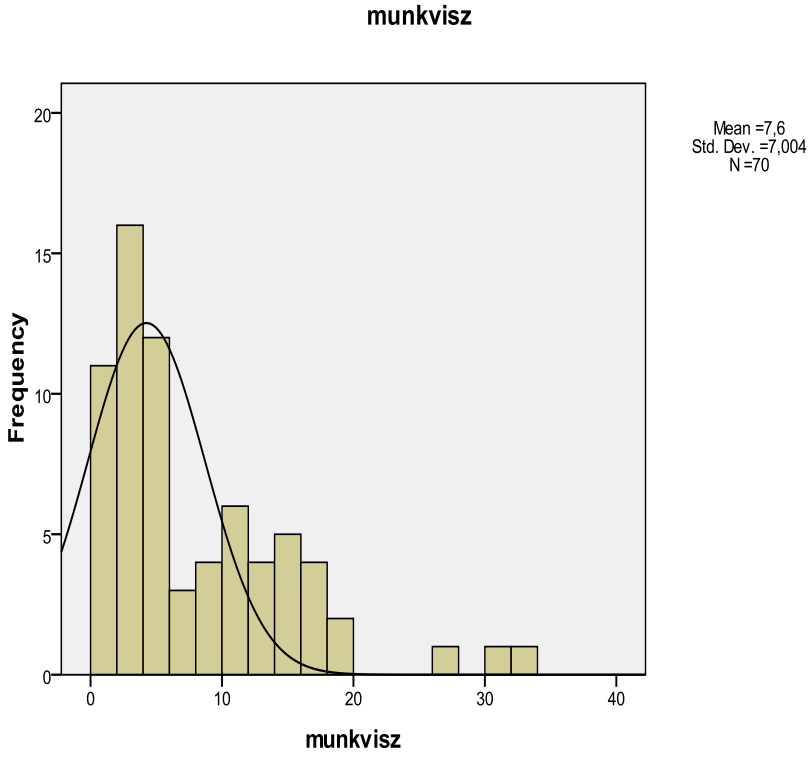
		Ekor	Munkvisz	jovedele
N	Valid	70	70	70
	Missing	0	0	0
Mean		42,40	7,60	29,07
Median		41,00	4,00	27,00
Mode		52	1	18[a]
Std. Deviation		11,946	7,004	18,715
Variance		142,707	49,055	350,241
Range		44	31	72
Minimum		19	1	3
Maximum		63	32	75

a. Multiple modes exist. The smallest value is shown

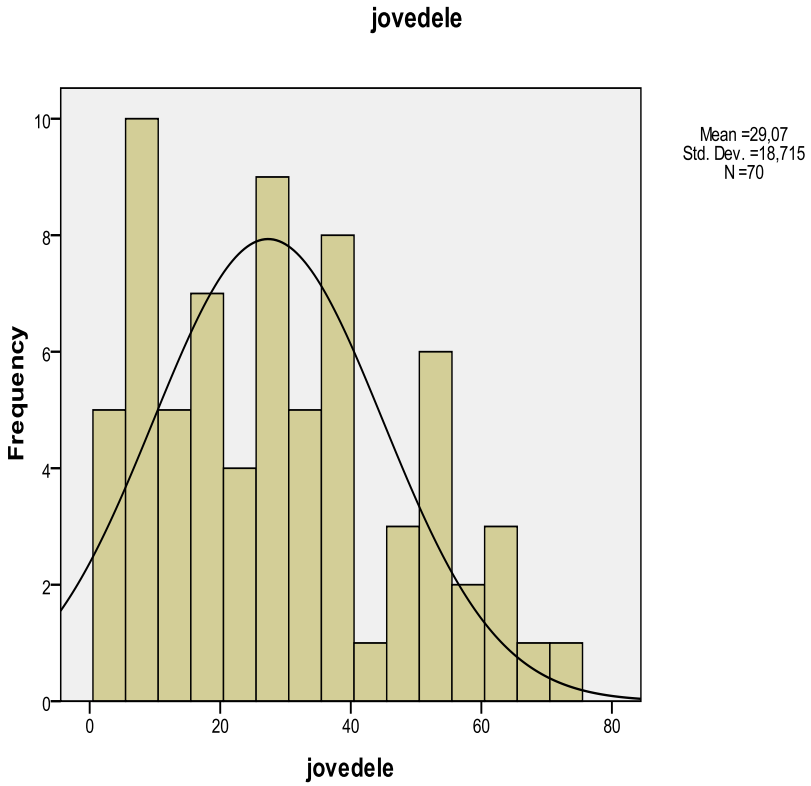
Folytonos változóinkat hisztogramon ábrázoljuk.



Bár az életkor alapvetően normális eloszlásúnak tűnik, ha jobban szemügyre vesszük a hisztogramot, akkor 2 csúcsú eloszlás ábrázolódik.



A munkaviszony egyértelműen ferde eloszlást mutat.



A jövedelem eloszlása is jól láthatóan ferde.

A hisztogramok alapján szerzett vizuális benyomást a normalitás statisztikai vizsgálatával lehet számszerűsíteni. Erre a Shapiro-Wilk tesztet használjuk. Ezt az Analyze → Destriptive Statistics → Explore menüponton belül érhetjük el.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
ekor	,136	70	,003	,952	70	,009
munkvisz	,225	70	,000	,834	70	,000
jovedele	,109	70	,040	,945	70	,004

a. Lilliefors Significance Correction

A táblázat utolsó oszlopában látható p értékek nem normális eloszlást igazolnak.

A nem normális eloszlást nagy valószínűséggel az okozza, hogy a folytonos változók bizonyos kategorikus változók alapján történő összehasonlításban különbséget mutatnak.

Ezért bontsuk mind a három vizsgált folytonos változót egyrészt a családi állapot, másrészt az adós statusa alapján alcsoportokra. Először nézzük meg a folytonos változók alcsoportjaira is a deskriptív statisztikát:

Descriptives

			Statistic	Std. Error
ekorcsalall1	Mean		34,00	3,420
	95% Confidence Interval for Mean	Lower Bound	26,11	
		Upper Bound	41,89	
	5% Trimmed Mean		34,06	
	Median		33,00	
	Variance		105,250	
	Std. Deviation		10,259	
	Minimum		19	
	Maximum		48	
	Range		29	
	Interquartile Range		19	
	Skewness		,057	,717
	Kurtosis		-,925	1,400
ekorcsalall2	Mean		54,44	2,346
	95% Confidence Interval for Mean	Lower Bound	49,03	
		Upper Bound	59,85	
	5% Trimmed Mean		54,94	
	Median		53,00	
	Variance		49,528	
	Std. Deviation		7,038	
	Minimum		39	
	Maximum		61	
	Range		22	
	Interquartile Range		9	
	Skewness		-1,304	,717
	Kurtosis		2,273	1,400
ekorcsalall3	Mean		39,89	2,679
	95% Confidence Interval for	Lower Bound	33,71	

	Mean	Upper Bound	46,07	
	5% Trimmed Mean		39,93	
	Median		41,00	
	Variance		64,611	
	Std. Deviation		8,038	
	Minimum		27	
	Maximum		52	
	Range		25	
	Interquartile Range		14	
	Skewness		-,159	,717
	Kurtosis		-,706	1,400
ekorstatus1	Mean		34,00	3,420
	95% Confidence Interval for Mean	Lower Bound	26,11	
		Upper Bound	41,89	
	5% Trimmed Mean		34,06	
	Median		33,00	
	Variance		105,250	
	Std. Deviation		10,259	
	Minimum		19	
	Maximum		48	
	Range		29	
	Interquartile Range		19	
	Skewness		,057	,717
	Kurtosis		-,925	1,400
ekorstatus2	Mean		47,56	3,709
	95% Confidence Interval for Mean	Lower Bound	39,00	
		Upper Bound	56,11	
	5% Trimmed Mean		47,56	
	Median		42,00	
	Variance		123,778	
	Std. Deviation		11,126	
	Minimum		32	
	Maximum		63	
	Range		31	

	Interquartile Range		20	
	Skewness		,224	,717
	Kurtosis		-1,627	1,400
ekorstatus3	Mean		35,89	3,323
	95% Confidence Interval for Mean	Lower Bound	28,23	
		Upper Bound	43,55	
	5% Trimmed Mean		35,60	
	Median		32,00	
	Variance		99,361	
	Std. Deviation		9,968	
	Minimum		24	
	Maximum		53	
	Range		29	
	Interquartile Range		18	
	Skewness		,549	,717
	Kurtosis		-,873	1,400
jovedelecsalall1	Mean		18,11	4,686
	95% Confidence Interval for Mean	Lower Bound	7,31	
		Upper Bound	28,92	
	5% Trimmed Mean		17,12	
	Median		12,00	
	Variance		197,611	
	Std. Deviation		14,057	
	Minimum		5	
	Maximum		49	
	Range		44	
	Interquartile Range		19	
	Skewness		1,471	,717
	Kurtosis		2,148	1,400
jovedelecsalall2	Mean		17,67	4,787
	95% Confidence Interval for Mean	Lower Bound	6,63	
		Upper Bound	28,71	
	5% Trimmed Mean		17,02	
	Median		12,00	

		Variance	206,250	
		Std. Deviation	14,361	
		Minimum	4	
		Maximum	43	
		Range	39	
		Interquartile Range	25	
		Skewness	,990	,717
		Kurtosis	-,402	1,400
jovedelecsalall3		Mean	47,56	4,385
	95% Confidence Interval for Mean	Lower Bound	37,44	
		Upper Bound	57,67	
		5% Trimmed Mean	47,67	
		Median	52,00	
		Variance	173,028	
		Std. Deviation	13,154	
		Minimum	28	
		Maximum	65	
		Range	37	
		Interquartile Range	24	
		Skewness	-,465	,717
		Kurtosis	-,988	1,400

Majd ezt követően alcsoportok szerint is vizsgáljuk meg a hisztogramok külön felvétele nélkül az eloszlás normalitását.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ekorcsalall1	,136	9	,200*	,949	9	,674
ekorcsalall2	,253	9	,100	,829	9	,044
ekorcsalall3	,172	9	,200*	,969	9	,885
ekorstatus1	,136	9	,200*	,949	9	,674
ekorstatus2	,247	9	,121	,901	9	,258
ekorstatus3	,207	9	,200*	,939	9	,572
jovedelecsalall1	,224	9	,200*	,848	9	,071
jovedelecsalall2	,209	9	,200*	,855	9	,085
jovedelecsalall3	,210	9	,200*	,911	9	,323

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Látszik, hogy ha alcsoportok szerinti bontásban vizsgáljuk az adatok normalitását, akkor lényegében minden esetben normális eloszlást kapunk, tehát az alcsoportonkénti összehasonlítások során, amennyiben a varianciák is homogének, parametrikus teszteket alkalmazhatunk.

5. Az elemzések elvégzése

1. A fentiek alapján az 1. pontban foglalt elemzés elvégzéséhez, amennyiben a varianciák homogének, úgy parametrikus statisztikai módszert tudunk használni. Jelen esetben 3 csoport összehasonlításához 1 utas varianciaanalízist szeretnénk használni: Analyze → Compare Means → One-way ANOVA.

Először az életkorbeli különbségeket vizsgáljuk a családi állapot szerint:

Test of Homogeneity of Variances

Ekor

Levene Statistic	Df1	df2	Sig.
,378	2	67	,687

A varianciák a Levene teszt alapján homogének, így nincs akadálya az ANOVA vizsgálat elvégzésének.

ANOVA

Ekor

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	1236,576	2	618,288	4,811	,011
Within Groups	8610,224	67	128,511		
Total	9846,800	69			

A táblázat utolsó oszlopában található p érték alapján a csoportok között szignifikáns különbség van, ezért elvégezzük a post hoc tesztet. Post hoc tesztként a legszigorúbb Bonferroni tesztet választottuk.

Post Hoc Tests

Multiple Comparisons

Ekor

Bonferroni

(I) csalall	(J) csalall	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
egyedülálló	elvált vagy özvegy	-12,143*	3,949	,009	-21,84	-2,45
	Házas	-7,356	3,392	,101	-15,69	,97
Elvált vagy özvegy	Egyedülálló	12,143*	3,949	,009	2,45	21,84
	Házas	4,787	3,322	,463	-3,37	12,94
házas	Egyedülálló	7,356	3,392	,101	-,97	15,69
	elvált vagy özvegy	-4,787	3,322	,463	-12,94	3,37

*. The mean difference is significant at the 0.05 level.

A fenti táblázat adatai és a korábban bemutatott deskriptív statisztikai adatok alapján elmondhatjuk, hogy az elvált, vagy özvegy egyének életkora (54,44 ± 7,04; átlag ± S.D.) szignifikánsan magasabb, mint az egyedülálló egyének életkora (34 ± 10,26).

Másodszor az életkorbeli különbségeket vizsgáljuk az adós statusa szerint:

Test of Homogeneity of Variances

Ekor

Levene Statistic	df1	df2	Sig.
,313	2	67	,732

A varianciák a Levene teszt alapján homogének, így nincs akadálya az ANOVA vizsgálat elvégzésének.

ANOVA

Ekor

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	475,566	2	237,783	1,700	,190
Within Groups	9371,234	67	139,869		
Total	9846,800	69			

A táblázat utolsó oszlopában található p érték alapján a csoportok között nincs szignifikáns különbség, ezért post hoc tesztet nem végzünk.

2. Elemzésünk 2. részében először azt szeretnénk megtudni, hogy életkor és a munkaviszony között van-e összefüggés. A fentiekben már megvizsgáltuk mindkét folytonos változó eloszlását, a Shapiro-Wilk teszt alapján mind a kettő nem normál eloszlást mutatott, ezért a nem parametrikus korrelációanalízist használunk: Analyze → Correlate → Bivariate.

Correlations

			ekor	munkvisz
Spearman's rho	ekor	Correlation Coefficient	1,000	,374**
		Sig. (2-tailed)	.	,001
		N	70	70
	munkvisz	Correlation Coefficient	,374**	1,000
		Sig. (2-tailed)	,001	.
		N	70	70

**. Correlation is significant at the 0.01 level (2-tailed).

A korrelációs koefficiens értéke (Spearman's rho) 0,374, a p érték szignifikáns különbséget mutat. Azaz az életkor és a munkaviszony között pozitív összefüggés van: minél nagyobb az életkor, annál hosszabb a munkaviszony.

Correlations

			ekor	jovedele
Spearman's rho	ekor	Correlation Coefficient	1,000	,082
		Sig. (2-tailed)	.	,500
		N	70	70
	jovedele	Correlation Coefficient	,082	1,000
		Sig. (2-tailed)	,500	.
		N	70	70

A korrelációs koefficiens értéke (Spearman's rho) 0,082, a p érték nem mutat szignifikáns különbséget. Azaz az életkor és a jövedelem között nincs összefüggés.

3. A korábban már elvégzett normalitásvizsgálat alapján az 3. pontban foglalt elemzés elvégzéséhez, amennyiben a varianciák homogének, úgy parametrikus statisztikai módszert tudunk használni. Jelen esetben is 3 csoport összehasonlításához 1 utas varianciaanalízist szeretnénk használni.

Test of Homogeneity of Variances

Jovedele

Levene Statistic	df1	df2	Sig.
2,238	2	67	,115

A varianciák a Levene teszt alapján homogének, így nincs akadálya az ANOVA vizsgálat elvégzésének.

ANOVA

jovedele

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7954,890	2	3977,445	16,438	,000
Within Groups	16211,753	67	241,966		
Total	24166,643	69			

A táblázat utolsó oszlopában található p érték alapján a csoportok között szignifikáns különbség van, ezért elvégezzük a post hoc tesztet. Post hoc tesztként itt is a legszigorúbb Bonferroni tesztet választottuk.

Post Hoc Tests

Multiple Comparisons

jovedele

Bonferroni

(I) csalall	(J) csalall	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
egyedülálló	elvált vagy özvegy	3,327	5,418	1,000	-9,98	16,63
	házas	-19,519*	4,654	,000	-30,95	-8,09
elvált vagy özvegy	egyedülálló	-3,327	5,418	1,000	-16,63	9,98
	házas	-22,846*	4,558	,000	-34,04	-11,65
házas	egyedülálló	19,519*	4,654	,000	8,09	30,95
	elvált vagy özvegy	22,846*	4,558	,000	11,65	34,04

*. The mean difference is significant at the 0.05 level.

A fenti táblázat adatai és a korábban bemutatott deskriptív statisztikai adatok alapján elmondhatjuk, hogy a házas egyének jövedelme (47,56 ± 13,15; átlag ± S.D.) szignifikánsan magasabb mind az elvált vagy özvegy egyének jövedelménél (17,67 ± 14,36), mind az egyedülálló egyének jövedelménél (18,11 ± 14,06), míg a két utóbbi között nincs szignifikáns különbség.

6. Konklúzió

Az elvégzett statisztikai elemzések alapján az alábbi következtetéseket vonhatjuk le. A mintavétel megfelelő voltát ellenőrizni kívánó vizsgálatok a várt eredményeket hozták. Egyrészt valószínűsíthető volt, hogy az elvált vagy özvegy egyének életkora magasabb az egyedülálló egyének életkoránál. Ezt a különbséget mintapopulációnkban is ki tudtuk mutatni. Másrészt az is valószínűsíthető volt, hogy az idősebb egyének hosszabb munkaviszonnyal rendelkeznek, ezt a különbséget is ki tudtuk mutatni mintapopulációnkban.
Az adós statusában nem találtunk különbséget, azaz a házasságban nem élőknek sem nehezebb kiszorítani a törlesztőrészletre szánt összeget, még akkor sem, ha ezen egyének jövedelme szignifikánsan kevesebb a házasságban élőknél. Azt vártuk volna, hogy életkor előre haladtával, mivel az emberek egyre hosszabb munkaviszonnyal rendelkeznek, valószínűleg egyre jövedelmezőbb pozícióba is kerülhetnek, de ilyen irányú összefüggést nem tudtunk kimutatni.