Többdimenziós statisztikai vizsgálatok pszichológia szakos hallgatók számára

Soltész-Várhelyi Klára

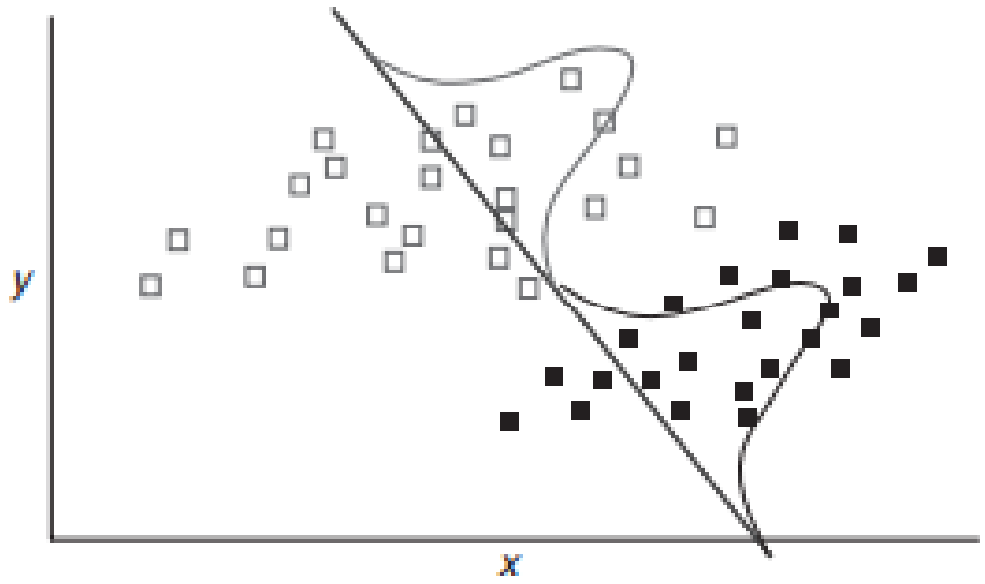
Diszkriminancia elemzés

Diszkriminancia analízis

- Célja olyan modell létrehozása, hogy néhány tulajdonság (folytonos prediktor változó) segítségével meg tudjuk jósolni, hogy az itemünk melyik (már létező) csoportba tartozik
- Klaszterelemzésre
  - Hasonlít annyiban, hogy esetek csoportosítására használható.
  - Különbözik, mert már meglévő csoportosítás alapján dolgozik. Az új elemeket szétválogatására használjuk
- Lineáris regresszió
  - Hasonlít, mert a prediktor változókból jósolja meg a kimeneti változót, a Y változót a prediktor változók súlyozott kombinációjaként számolja
  - Különbözik, mert míg a lineáris regresszió feltételezi, hogy a függő változó folytonos, a logisztikus regresszióban pedig dichotóm, a DA-ban kategoriális függő változóval dolgozunk

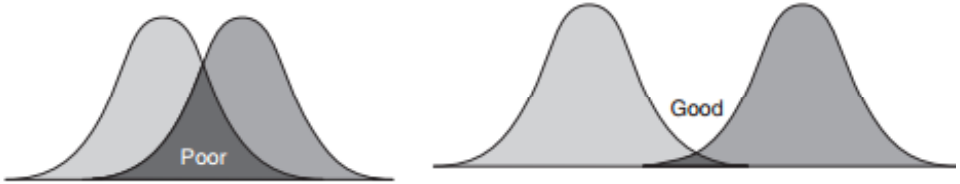
Diszkriminancia analízis

- Egyenlete:
  - $D= v_1*X_1 + v_2*X_2 + v_3*X_3 + ... + v_i*X_i + a$
  - D=diszkrimináló funkció (diszkriminancia érték/discriminant score), v=súlyozás (diszkriminancia-koefficiens) , X=prediktor változóra adott érték, 1-i=hányadik prediktor változóról van szó, a=hiba
  - Cél olyan egyenlet létrehozása, melyben a D kategóriái minél jobban elkülönülnek egymástól

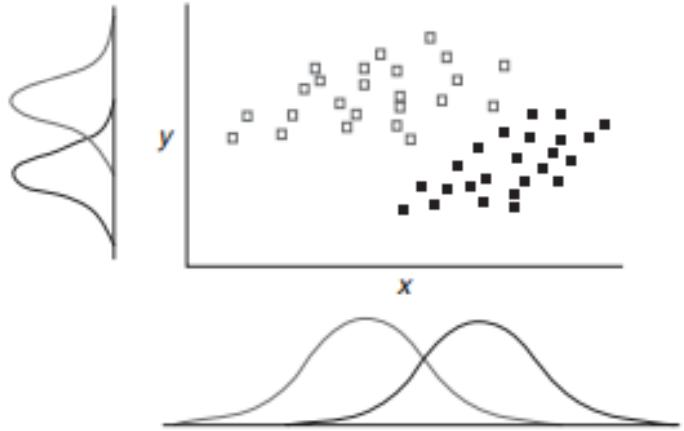


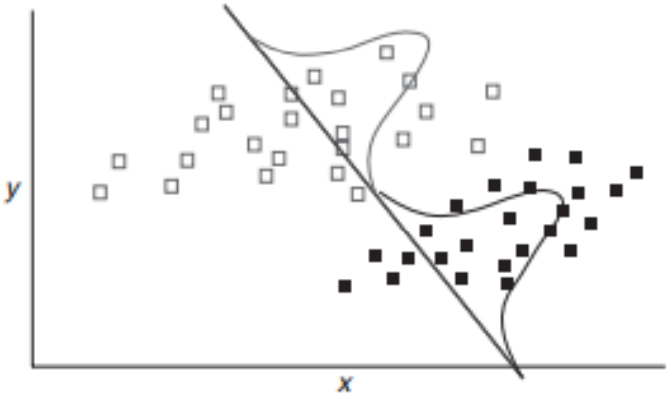
Diszkriminancia analízis

- A DA súlyozza a prediktor változókat, és a kombinációjukból létrehoz egy közös, új változót, a diszkriminancia értéket, mely mentén minden csoport elkülönül és saját normális eloszlással rendelkezik.
- Mit jelent az, hogy két csoport elkülönül?



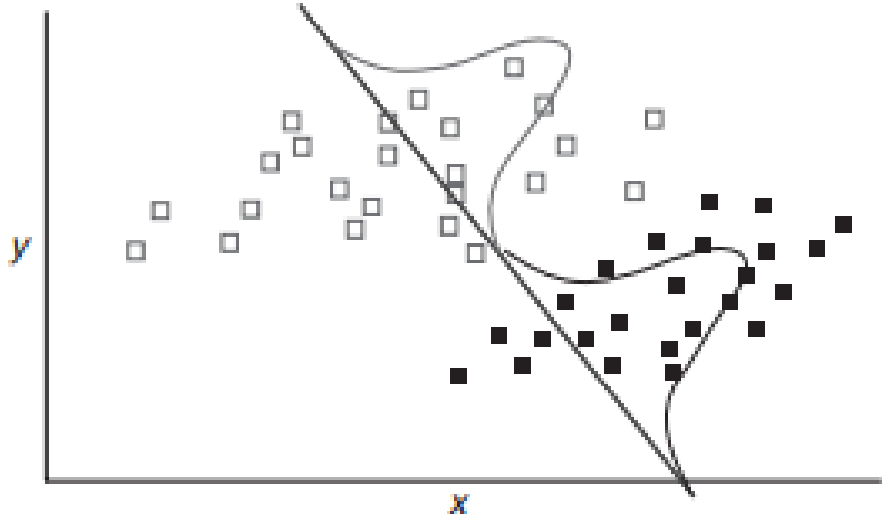
- Hogy működik a DA?
- Sem az x sem az y változó mentén nem különül a két csoport jól el (bár y mentén jobban elkülönül, az y-nak nagyobb a diszkrimináló ereje).
- Az új tengely leírható az x és y kombinációjaként. Az új tengely mentén elkülönül a két csoportunk.
- Mindig csoport-1 tengelyt keres





Diszkriminancia analízis

- Ha megvan a diszkriminancia érték egyenese, meghatározunk rajta egy cut-off pontot, azokban az esetekben, ahol a D értéke a cut-off pontnál kisebb, az egyik csoportba, ahol nagyobb, a másik csoportba predikáljuk az egyedet.
- A cut-off pont a két csoport közepéhez tartozó D érték átlaga.

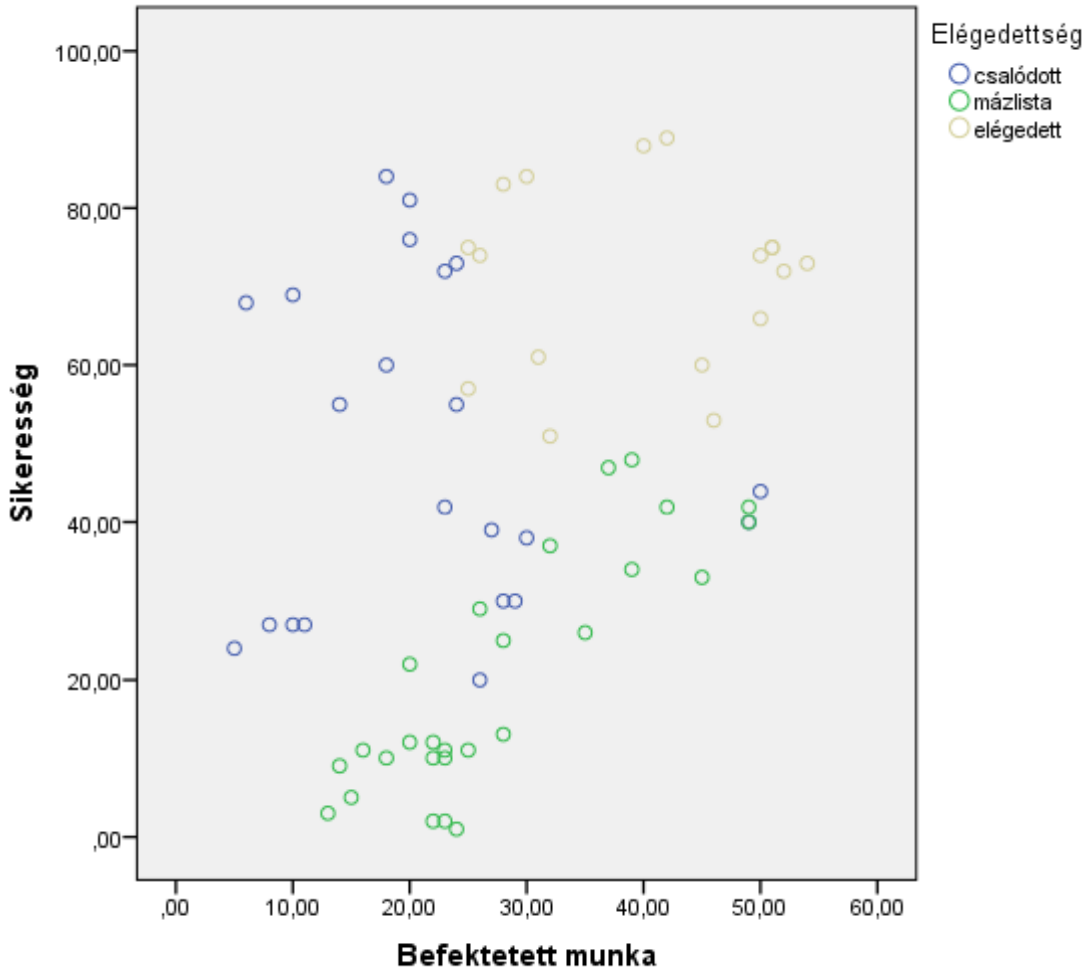


Diszkriminancia analízis feltételei

- Az esetek függetlenek egymástól
- A prediktor változók normál eloszlásúak
- Legalább két csoport van, és minden eset egy és csakis egy csoportba tartozik.
- A csoportok természetesek. Egy skálát szétszedni alacsony/közepes/magas értékű csoportra nem természetes csoportosítás (az utóbbi csak akkor használható csoportosító változóként, ha a skálán a kívánt csoportok között jól látható hézagok vannak)
- A csoportok az adatgyűjtés előtt kerülnek kialakításra
- Olyan prediktor változókat válasszunk, melyek önmagukban is képesek valamelyest a csoportokat szétválasztani
- A csoportok elemszáma nagyjából azonos legyen és minden csoportban legalább ötször annyi ember legyen, mint amennyi prediktor változó van

Nézzünk rá!

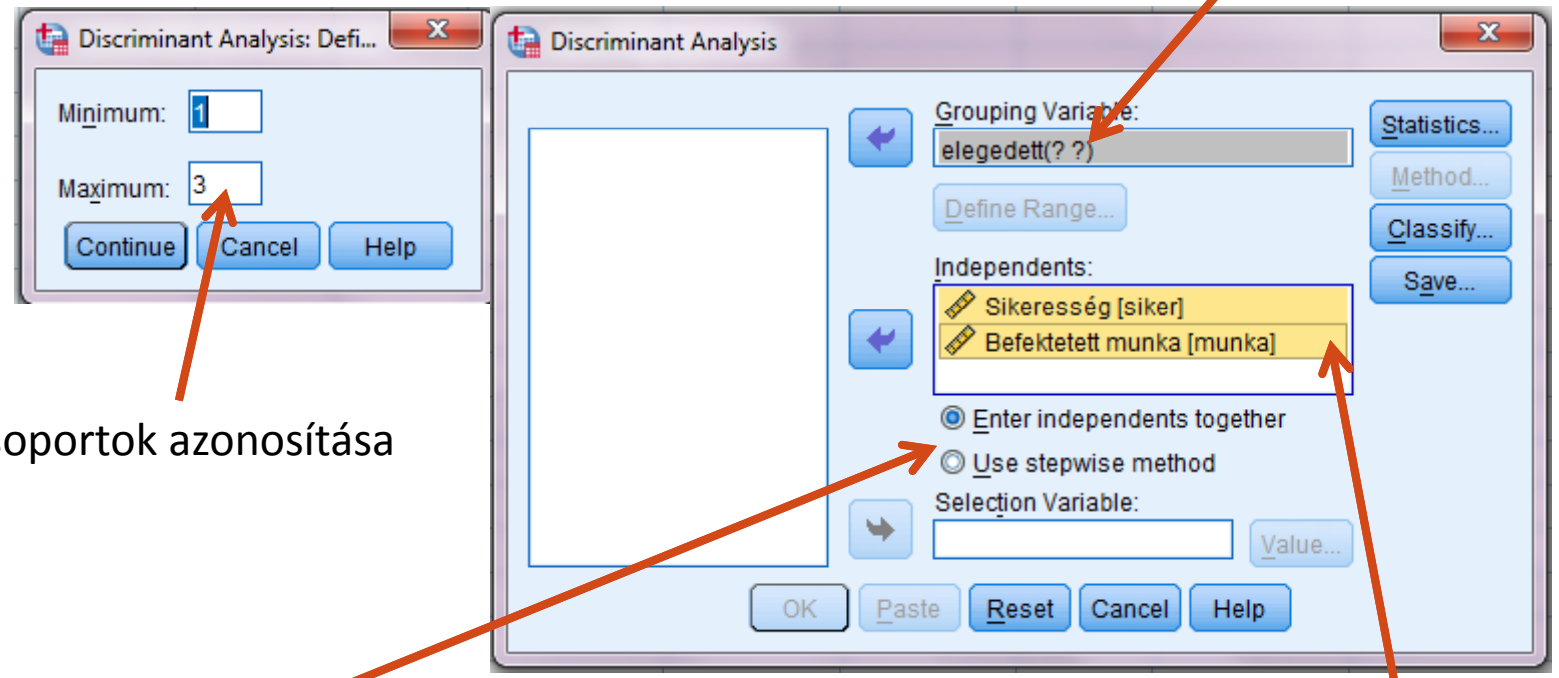
- Adatfájl: tobbval08_diszkrim_elegedettseg.sav
- Graph / Legacy Dialogs / Scatter Dot



SPSS-ben

- Analyze / Classify / Discriminant

Csoportosító változó

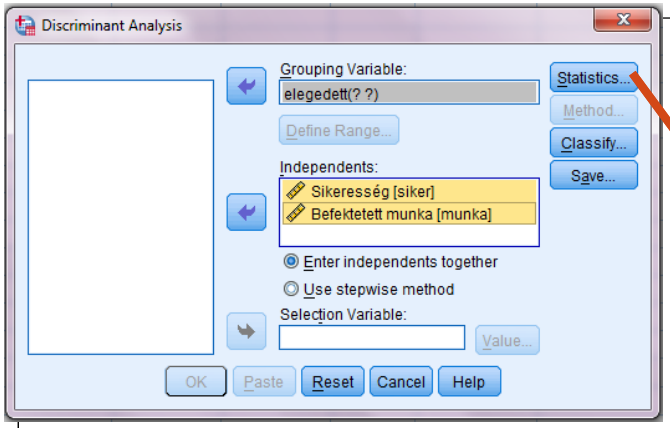


Csoportok azonosítása

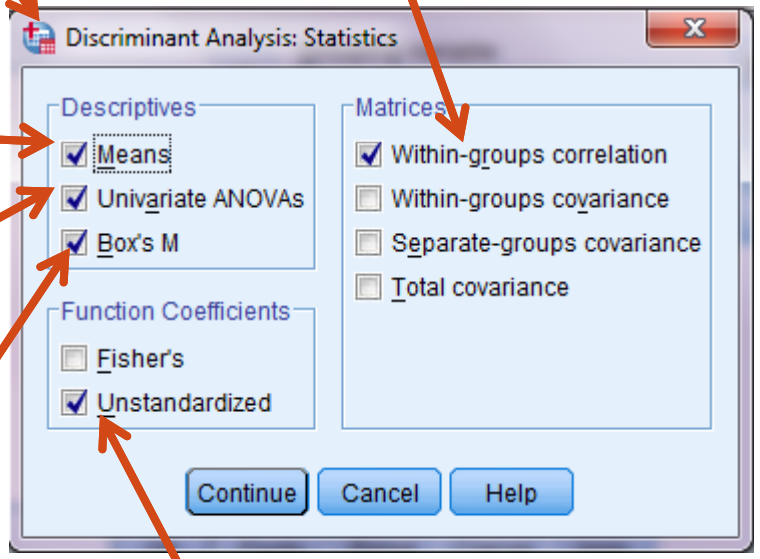
A FA-hoz hasonlóan a változók modellbeléptetési sorrendje adható meg:
Enter: egyszerre a modellbe kényszeríti az összes változót
Stepwise: először a legtöbbet magyarázó kerül a modellbe

Prediktor változók

SPSS-ben



Korrelációs tábla, multikollinearitásra ellenőrzünk vele



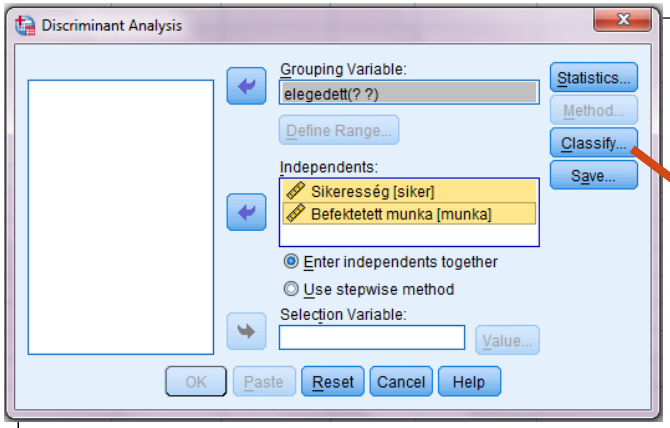
Leíró statisztikák

ANOVA vizsgálat: van a prediktor változók értékében különbség a csoportok között. Van-e esély arra, hogy ezek a változók segítségével szétválasszuk a csoportjainkat?

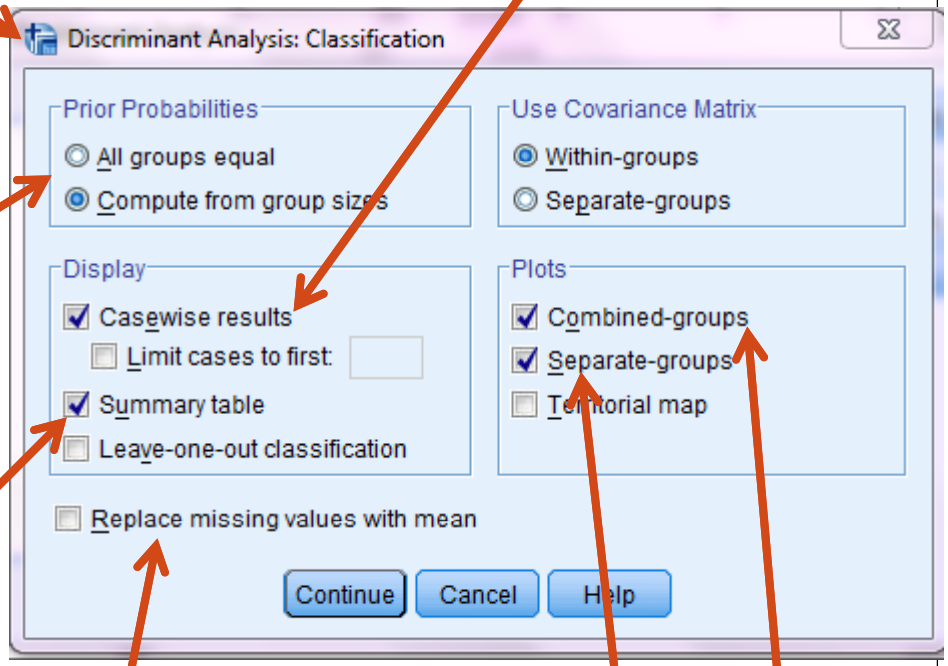
Box's M: DA feltétele a kovarianciák homogenitása.

A változók fontosságáról ad információt, hasonlóan mint regresszióelemzésben a regressziós együttható.
A standardizáltat (regresszióban ez a B) mindenképp megkapjuk, de itt kérhetjük ki a standardizálatlant is (regresszióban b1)

SPSS-ben



Casewise result: gyanús esetek után keresünk vele (regresszióelemzéshez hasonlóan)

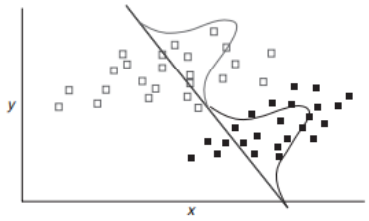


Ha valamelyik csoportból több van a valóságban, akkor az abba a csoportba való kerülés is valószínűbb.
A csoportba való kerülés kezdeti valószínűsége legyen egyforma minden csoportra, vagy az, hogy melyik csoportunkban hány eset van, követi azt, a valóságban az emberek mekkora része tartozik a különböző csoportokba

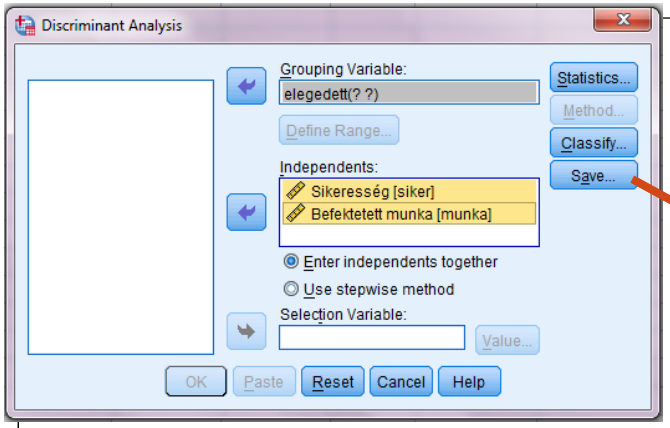
Arról információ, hogy hány esetet sikerült helyesen/helytelenül besorolnia

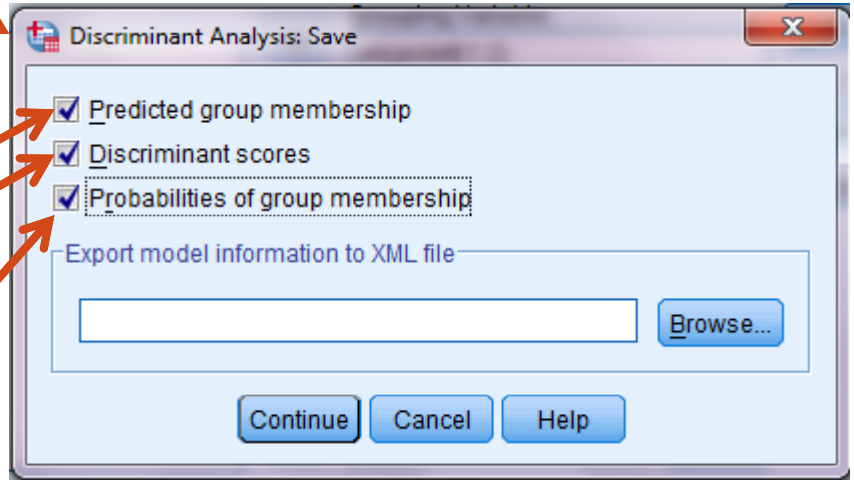
Ha van hiányzó érték, az átlaggal helyettesíthetjük

A diszkrimináns funkció eloszlása csoportok szerint



SPSS-ben





Mit jósol a DA, melyik csoportba tartozik az eset

A diszkrimináns értéke

Mekkora a valószínűsége, hogy a csoportokba tartozik

Output

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		66	100,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Total	0	,0
Total		66	100,0

Információs tábla arról, hogy van-e hiányzó, hibás érték, elemzésbe be nem vett eset

Group Statistics

Elégedettség		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
mázlista	Sikeresség	49,1364	20,76236	22	22,000
	Befektetett munka	21,5000	11,87534	22	22,000
csalódott	Sikeresség	20,2593	15,24566	27	27,000
	Befektetett munka	27,7407	10,61056	27	27,000
elégedett	Sikeresség	71,1765	11,62559	17	17,000
	Befektetett munka	39,8824	10,84483	17	17,000
Total	Sikeresség	43,0000	26,47088	66	66,000
	Befektetett munka	28,7879	13,04491	66	66,000

Leíró statisztikák

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Sikeresség	,379	51,631	2	63	,000
Befektetett munka	,703	13,339	2	63	,000

ANOVA vizsgálat: van a prediktor változók értékében különbség a csoportok között. Van-e esély arra, hogy ezek a változók segítségével szétválasszuk a csoportjainkat? Ha mind nagyon gyenge, akkor nincs sok esélyünk arra, hogy DA sikeres lesz

Pooled Within-Groups Matrices

		Sikeresség	Befektetett munka
Correlation	Sikeresség	1,000	,274
	Befektetett munka	,274	1,000

Korrelációs tábla – multikollinearitásra (r>.9) keresünk

Output

Box's Test of Equality of Covariance Matrices

Log Determinants

Elégedettség	Rank	Log Determinant
mázlista	2	11,007
csalódott	2	8,892
elégedett	2	9,673
Pooled within-groups	2	10,350

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Log determináns értéke

Box’s test

Test Results

Box's M		34,913
F	Approx.	5,539
	df1	6
	df2	46915,727
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

A DA feltétele az, hogy a csoportosító változó által létrehozott csoportok nem különböznek a kovariancia mátrixukban.
A Box’s test azt teszteli, hogy eltérnek-e a csoportok, ha szignifikáns, megsérül e kritérium. Ekkor azokat a pcsoportokat, melyek túl alacsony log determinánssal rendelkeznek, ki kell venni az elemzésből
Nagy elemszám esetén a Box’s test nagy valószínűséggel szignifikáns, de ekkor nem kell figyelembe venni

Output

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1,639[a]	83,0	83,0	,788
2	,336[a]	17,0	100,0	,502

a. First 2 canonical discriminant functions were used in the analysis.

Az Eigenvalues tábla a diszkrimináló funkcióink „jóságáról” ad információt
Mindig maximum csoport-1 funkció van: Nekünk 3 csoportunk volt, ezért 2 funkciónk keletkezett

Értelmezése hasonló a faktoranalízishez
Mennyire jól különíti el a funkció a csoportokat

Diszkrimináló funkció sorszáma

Megadja, hogy a diszkrimináló funkció mennyire korrelál a prediktor változókkal. A szám négyzete megadja, hogy az adott diszkrimináló funkció mennyit tudott megmagyarázni a prediktor változó varianciájából. Tehát most például az első funkció R^2= 0.788^2=.62 tehát 62%-át magyarázza a varianciának

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,284	78,774	4	,000
2	,748	18,114	1	,000

Megadja, hogy a diszkrimináló funkció magyarázóereje szignifikáns-e.

A Wilks’ Lambda a diszkrimináló funkció teljes varianciájának csoportosító változó által meg **nem** magyarázott hányadát jelzi. Ha csak egy diszkrimináló funkciónk lenne, akkor ez pont a canonical correlation négyzetének ellentettje lenne. Az R2-tel ellentétben itt a minél kisebb szám az előnyös

Output

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Sikeresség	1,004	-,269
Befektetett munka	-,017	1,040

Standardizált diszkriminancia együttható (súly) – a különböző prediktor változók milyen mértékben vesznek részt az adott diszkrimináló funkcióban. Itt például az első diszkrimináló funkció kialakulásában nagyobb mértékben vesz rész a sikeresség – mint a regresszió analízisben a B volt

Structure Matrix

	Function	
	1	2
Sikeresség	1,000*	,016
Befektetett munka	,259	,966*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

Másik mutató a prediktor változók és diszkrimináló funkció kapcsolatára: mekkora a prediktor változók és a funkciók között a Pearson korreláció

Canonical Discriminant Function Coefficients

	Function	
	1	2
Sikeresség	,061	-,016
Befektetett munka	-,002	,094
(Constant)	-2,566	-1,996

Unstandardized coefficients

A nem standardizált diszkriminancia együttható. Mint a regresszió elemzésben a b_0 és b_1 volt
E segítségével felállítható a diszkrimináló funkció kiszámítására az egyenlet. Itt például:
D1 = -2,566 + 0.061*sikeresség + (-0.002)*befektetett munka

Functions at Group Centroids

Elégedettség	Function	
	1	2
mázlista	,383	-,782
csalódott	-1,379	,272
elégedett	1,693	,581

Unstandardized canonical discriminant functions evaluated at group means

A csoportok közepéhez milyen diszkrimináló funkció érték tartozik (részletesebben a következő dián)

Output

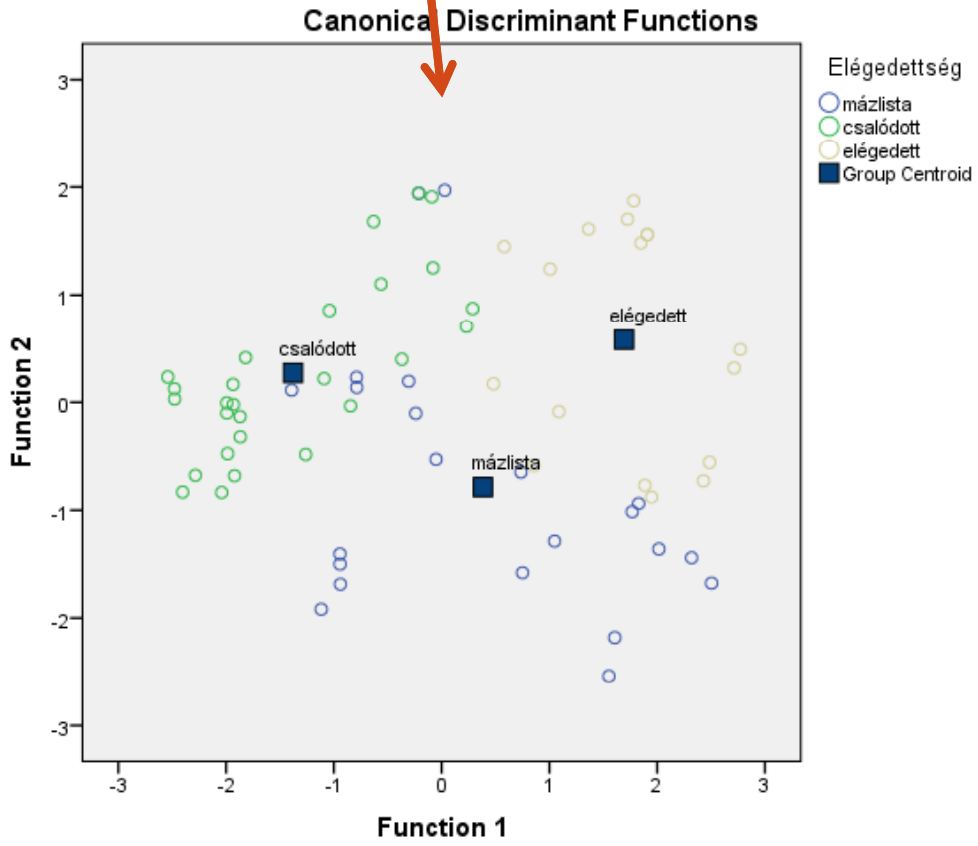
Functions at Group Centroids

Elégedettség	Function 1	Function 2
mázlista	,383	-,782
csalódott	-1,379	,272
elégedett	1,693	,581

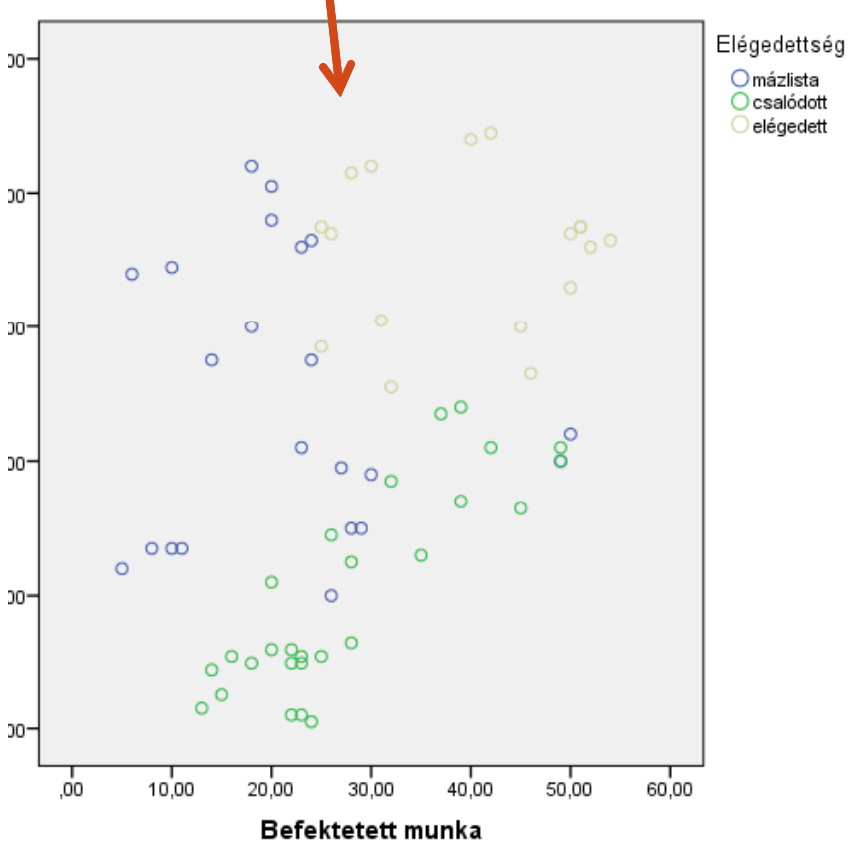
Unstandardized canonical discriminant functions evaluated at group means

A csoportok közepéhez milyen diszkrimináló funkció érték tartozik

X és Y tengelyen a két diszkrimináló funkciónk



X és Y tengelyen a két eredeti prediktor változónk



Output

Classification Statistics

Classification Processing Summary

Processed		66
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		66

Volt-e valami, ami miatt ki kellett hagyni esetet az elemzésből

Prior Probabilities for Groups

Elégedettség	Prior	Cases Used in Analysis	
		Unweighted	Weighted
mázlista	,333	22	22,000
csalódott	,409	27	27,000
elégedett	,258	17	17,000
Total	1,000	66	66,000

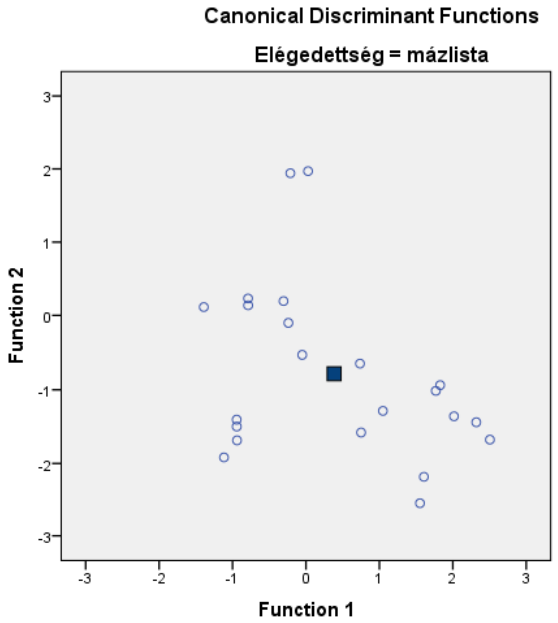
Mennyi az előzetes valószínűsége egy-egy csoportba kerülésnek. Az eltérő csoport-elemszámnál beszéltünk róla

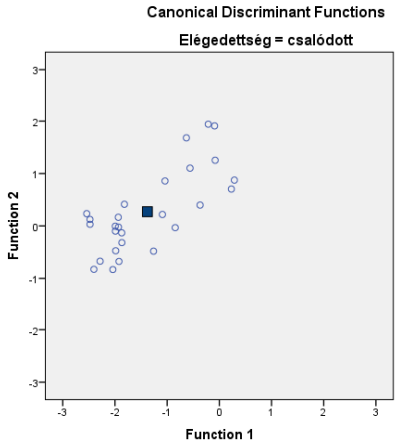
Casewise Statistics

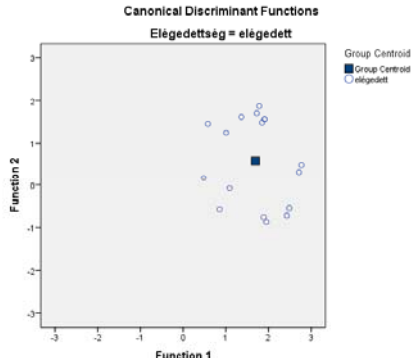
	Case Number	Actual Group	Highest Group				
			Predicted Group	P(D>d | G=g)		P(G=g | D=d)	Squared Mahalanobis Distance to Centroid
				p	df		
Original	1	1	2**	,126	2	,687	4,146
	2	1	1	,348	2	,586	2,110
	3	1	1	,653	2	,485	,851
	4	1	1	,277	2	,626	2,567
	5	1	1	,170	2	,611	3,542
	6	1	1	,071	2	,618	5,300
	7	3	1**	,632	2	,531	,917
	8	1	1	,108	2	,946	4,455
	9	3	3	,658	2	,945	,837
	10	3	1**	,324	2	,511	2,257

Minden egyes esetről infó
Melyik csoportba tartozik.
A DA hova sorolta és milyen valószínűséggel. Csillaggal jelölve minden hibás besorolást.
Milyen távol van a csoportja középpontjától

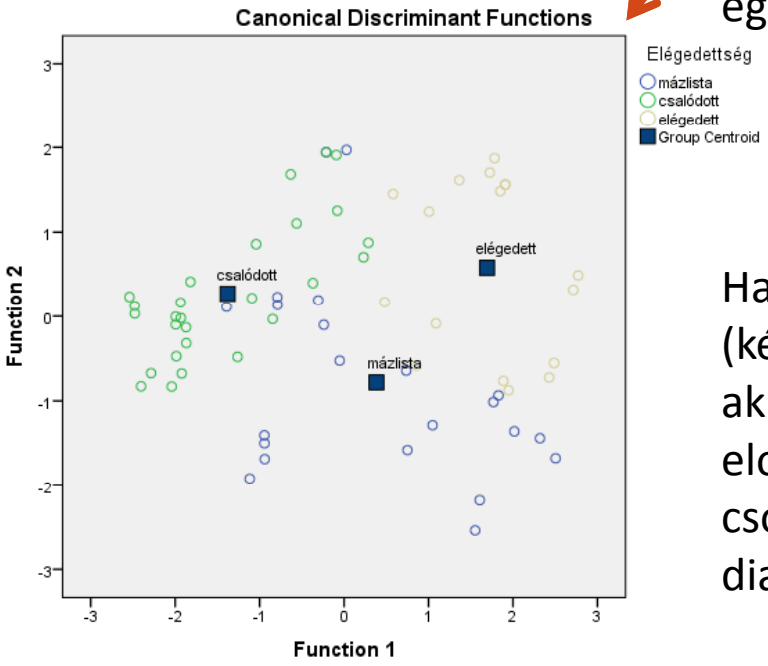
Output

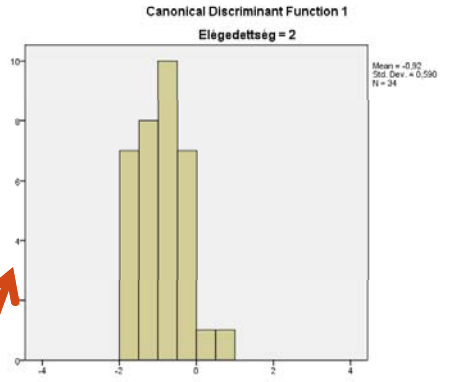




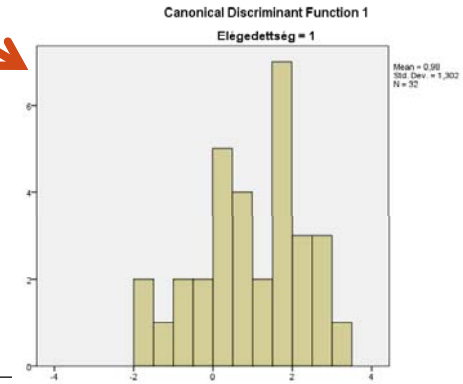


A csoportok és középpontjuk elhelyezkedése a DA funkció által kifeszített térben egyenként és együtt





Ha csak egy DA funkciónk lenne (két csoport esetén van ez így), akkor scatter plot helyett eloszlásgörbét mutat csoportonként, és közös diagram pedig nincs



Output

Classification Results[a]

		Elégedettség	Predicted Group Membership			Total
			mázlista	csalódott	elégedett	
Original	Count	mázlista	16	6	0	22
		csalódott	1	25	1	27
		elégedett	5	0	12	17
	%	mázlista	72,7	27,3	,0	100,0
		csalódott	3,7	92,6	3,7	100,0
		elégedett	29,4	,0	70,6	100,0

a. 80,3% of original grouped cases correctly classified.

Csoportosítás: mit hova sorolt
Megértéshez elkezdem értelmezni:
Itt 16 esetet helyesen mázlistának sorolt (ez a mázlisták 72,7%-a. 6 esetet a mázlisták közül csalódottnak sorolt, ez az esetek(27,3%-a), végül egy mázlistát sem sorolt elégedettnek. És így tovább.

A lábjegyzetből kiolvashatjuk, hogy az esetek 80,3%-át csoportosította helyesen.

Output

Prediktor változók

Eredeti csoportosító változó

DA hova sorolta

Értéke az első DA funkción

Értéke a második DA funkción

Annak valószínűsége, hogy az első/második/harmadik csoportba tartozik

siker	munka	elegedett	Dis_1	Dis1_1	Dis2_1	Dis1_2	Dis2_2	Dis3_2
40,00	49,00	1	2	-,21274	1,94113	,09137	,68702	,22161
73,00	24,00	1	1	1,82797	-,93605	,58605	,00583	,40812
39,00	27,00	1	1	-,24005	-,10235	,48534	,44452	,07014
27,00	8,00	1	1	-,93950	-1,68610	,62554	,37028	,00418
24,00	5,00	1	1	-1,11702	-1,91819	,61066	,38698	,00236
84,00	18,00	1	1	2,50467	-1,67665	,61834	,00086	,38080

És ha most van néhány új, csoportosítatlan esetem,

30,00	29,00	1	2	-,78929	,23123	,22040	,75516	,02444
27,00	11,00	1	1	-,94405	-1,40523	,55189	,44274	,00537
70,00	30,00	.	.	.	.	.	.	.
20,00	30,00	.	.	.	.	.	.	.

akkor lefuttatom az elemzést, és kapok a csoportjukra egy predikciót. Ezért csináltuk az egészet☺

30,00	29,00	1	2	-,78929	,23123	,22040	,75516	,02444
27,00	11,00	1	1	-,94405	-1,40523	,55189	,44274	,00537
70,00	30,00	.	3	1,63680	-,32552	,44010	,01167	,54823
20,00	30,00	.	2	-1,39771	,48745	,07051	,92450	,00500