TÖBBVÁLTOZÓS REGRESSZIÓ ÉS TÖBBVÁLTOZÓS MÓDSZEREK

Boda Krisztina
SZTE Orvosi Fizikai és Orvosi Informatikai Intézet

Szeged, 2016.

Regressziós modellek

- Regressziós modellek: egy függő változó és egy vagy több független (magyarázó) változó közötti kapcsolat.
- A cél a jelenséget jól leíró, de azért takarékos modell keresése (elegendő számú változó legyen benne, de ne túl sok)

Regressziós modellek

- <u>Egyszerű lineáris regresszió</u>: egy folytonos függő és egy folytonos független változó $y=b_0+b_1x$.
- <u>Többszörös lineáris regresszió</u>: egy folytonos függő és több folytonos független változó $y=b_0+b_1x+b_2x+\ldots+b_px_p$
- Bináris logisztikus regresszió: egy bináris függő változó és egy vagy több kategorikus vagy folytonos független változó

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}}$$

Többszörös lineáris regresszió

- Egyenlete:
  $y=b_0+b_1x+b_2x+\ldots+b_px_p$ (p a koordináták száma)
- Grafikonja: hipersík a p+1 dimenziós térben
- Együtthatók becslése: legkisebb négyzetek módszere
- Hipotézisvizsgálatok
  - A modell egészének vizsgálata.
    H0: minden együttható elméleti értéke 0, azaz y egyik xi-től sem függ
    tesztelése F(p, n-p-1) statisztikával
  - Az egyes együtthatók vizsgálata
    H0: az adott együttható elméleti értéke 0
    tesztelése t-eloszlású statisztikával, df=n-p-1
  - Feltétel: a reziduálok normális eloszlásúak

Korrelációk

- Többszörös korrelációs együttható, R: a függő változó és a magyarázó változók együttese közötti korreláció (az illeszkedés jóságát, kapcsolat erősségét jellemzi)
- Többszörös determinációs együttható, R^2: a függő változó varianciájának hány százaléka magyarázható a független változókkal.
- Kétváltozós korrelációk (Pearson korrelációk)
- Parciális korrelációk: két független változó közötti korreláció, miután a többi változó lineárishatását kiszűrtük.

Példa (orvkerd2015). Hogyan függ a jelenlegi testtömeg a magasságtól és az 1 évvel ezelőtti testtömegtől?

- R-output

```
Call:
lm(formula = tomeg ~ magassag + tomeg1, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.2601  -1.7225  -0.0147   1.7530  10.9844 

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)    
(Intercept) -20.57070    4.79587  -4.289 0.00002844 ***
magassag      0.16892    0.03350   5.042 0.00000107 ***
tomeg1        0.87377    0.02622  33.322    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.138 on 191 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.9303,	Adjusted R-squared:  0.9296 
F-statistic:  1276 on 2 and 191 DF,  p-value: < 2.2e-16
```

Regresszós egyenlet:
tömeg=-20.57+0.169mag+0.89tomeg1

Mindegyik együttható szignifikán

Többszörös R2=0.93, a tömeg összvarianciájának 93%-a magyarázható a független változóktól való lineáris függésével. A modell is szignifikáns (F(2,191)=1276, p<0.0001

SPSS output

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,965[a]	,930	,930	3,138

a. Predictors: (Constant), tomeg1, magassag

b. Dependent Variable: tomeg

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25128,262	2	12564,131	1275,615	,000[b]
	Residual	1881,248	191	9,849		
	Total	27009,510	193			

a. Dependent Variable: tomeg

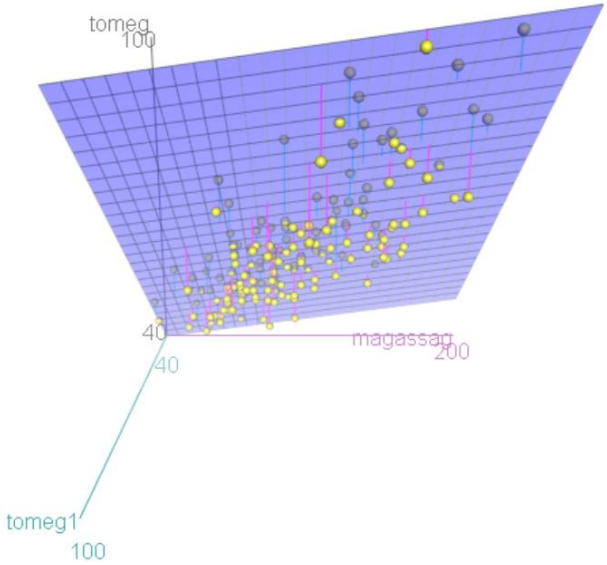
b. Predictors: (Constant), tomeg1, magassag

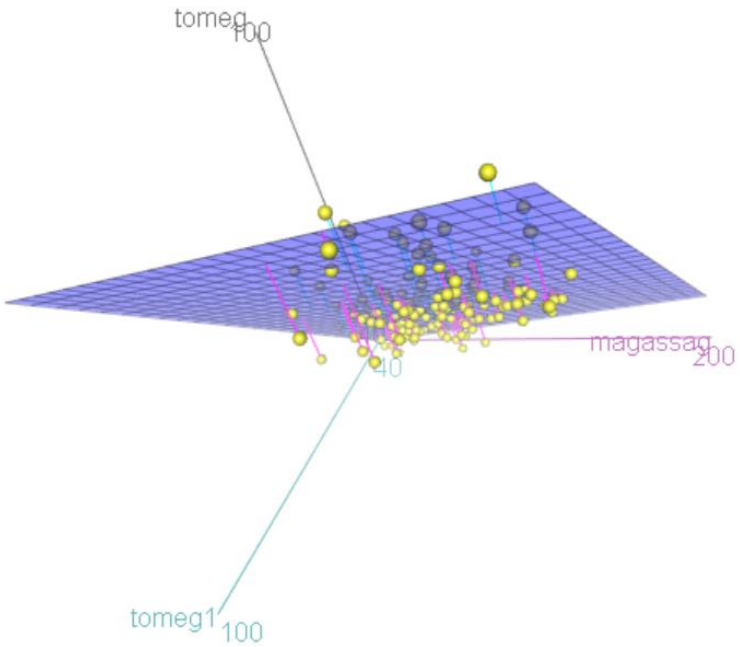
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-20,571	4,796		-4,289	,000					
	magassag	,169	,034	,132	5,042	,000	,725	,343	,096	,535	1,869
	tomeg1	,874	,026	,870	33,322	,000	,960	,924	,636	,535	1,869

a. Dependent Variable: tomeg

3D ábra Rcommanderrel





Multikollinearitás

- Akkor fordul elő, ha a magyarázó változók erősen korreláltak.
- Következmény: lehet, hogy külön-külön mindegyik szoros lineráris kapcsolatban van a függő változóval, de ha együttesen szerepelnek a modellben, egymást rontva nemszignifikánssá válnak.

- A szisztolés vérnyomás függ-e a koleszterinszinttől?

```
lm(formula = sbp ~ összkoleszterin, data = Dataset)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     104.8777     2.4713   42.44   <2e-16 ***
összkoleszterin   4.1009     0.4623    8.87   <2e-16 ***
```

Függ a koleszterinszinttől ($p < 0.001$)

- Függ-e a szisztolés vérnyomás a diasztoléstól, az életkortól, a koleszterinszinttől és a vércukorszinttől?

```
> RegModel.7 <- lm(sbp~dbp+életkor+összkoleszterin+vércukor, data=Dataset)
> summary(RegModel.7)
Call:
lm(formula = sbp ~ dbp + életkor + összkoleszterin + vércukor,     data =
Dataset)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     12.38731    2.63229   4.706 2.75e-06 ***
dbp              1.13904    0.02970  38.356  < 2e-16 ***
életkor          0.36156    0.03457  10.458  < 2e-16 ***
összkoleszterin  0.09698    0.32884   0.295    0.768
vércukor         1.38075    0.25423   5.431 6.49e-08 ***
---

Residual standard error: 12.51 on 1551 degrees of freedom
  (79 observations deleted due to missingness)
Multiple R-squared:  0.5874,    Adjusted R-squared:  0.5864
F-statistic: 552.1 on 4 and 1551 DF,  p-value: < 2.2e-16
```

Nincs szignifikáns függés a koleszterinszinttől ($p = 0.768$)

Az eltérés magyarázta a magyarázó változók egymással is korrelálnak! (NEM FÜGGETLENEK!!!)

	dbp	életkor	összkoleszterin	sbp	vércukor
dbp	1.0000000	0.2620757	0.1935321	0.7330061	0.2282176
életkor	0.2620757	1.0000000	0.3490340	0.3902484	0.2718843
összkoleszterin	0.1935321	0.3490340	1.0000000	0.2163103	0.1813650
sbp	0.7330061	0.3902484	0.2163103	1.0000000	0.2970396
vércukor	0.2282176	0.2718843	0.1813650	0.2970396	1.0000000

Mit tehetünk?

- az érintett változókat megpróbáljuk kihagyni a modellből, de ez ne menjen a magyarázó erő rovására
- származtatott változókkal dolgozunk

Feltételek ellenőrzése

- Reziduálok normalitása az egyszerű lineáris regresszióhoz hasonlóan tesztelhető
- Multikollinearitás tesztelése:
  - Korrelációs ábrák
  - Tolerancia (0-hoz közeli) vagy VIF (nagy érték) számítása
- Több változó közötti választás – modellválasztás:
  - Best subset regresszió
  - Lépésenkénti (stepwise) módszerek – az automatizmus veszélyes is lehet! Figyeljünk a kimaradó változókra

Többváltozós módszerekről általában

- A többváltozós módszereket akkor alkalmazzuk, ha a kiinduló adathalmazunk olyan, hogy az objektumokról egyszerre több tulajdonságot is megmértünk. Az így keletkező statisztikai minta nem egy számsor, hanem számok táblázata, amelyben a sorok az egy egyedről mért különböző tulajdonságokat, az oszlopok a különböző egyedekről mért azonos jellemzőt tartalmazzák. Statisztikai nyelven a vizsgált változó ún. vektorváltozó.

Az adatok vizsgálata, feltáró elemzés

- A változók többnyire korrelálnak (enélkül nincs is sok értelme a többváltozós módszereknek) -> cél: egyszerűsítés,pl. dimenziócsökkentés (feltáró jellegű elemzés)
- A többváltozós módszerek figyelembe veszik az adatok közötti korrelációt, és nem növelik az első fajta hibát

Mely változók korrelálnak legjobban?

```
> ###############mik korrelálnak legjobban?#########
> #http://little-book-of-r-for-multivariate-
analysis.readthedocs.org/en/latest/src/multivariateanalysis.html
>
> mosthighlycorrelated <- function(mydataframe,numtoreport)
+ {
+   # find the correlations
+   cormatrix <- cor(mydataframe)
+   # set the correlations on the diagonal or lower triangle to zero,
+   # so they will not be reported as the highest ones:
+   diag(cormatrix) <- 0
+   cormatrix[lower.tri(cormatrix)] <- 0
+   # flatten the matrix into a dataframe for easy sorting
+   fm <- as.data.frame(as.table(cormatrix))
+   # assign human-friendly names
+   names(fm) <- c("First.Variable", "Second.Variable","Correlation")
+   # sort and print the top n correlations
+   head(fm[order(abs(fm$Correlation),decreasing=T),],n=numtoreport)
+ }
> mosthighlycorrelated(ha[3:19], 15)
    First.Variable Second.Variable Correlation
72           tomeg          tomeg1   0.8928101
90          tomeg1      Kgszeretne   0.8433316
105       magassag      cmszeretne   0.8380684
89           tomeg      Kgszeretne   0.8057002
88        magassag      Kgszeretne   0.7929413
103            nem      cmszeretne  -0.7797175
86             nem      Kgszeretne  -0.7735859
108     Kgszeretne      cmszeretne   0.7630942
35             nem        magassag  -0.6967270
71        magassag          tomeg1   0.6889971
54        magassag           tomeg   0.6615547
69             nem          tomeg1  -0.6132826
106          tomeg      cmszeretne   0.6082726
107         tomeg1      cmszeretne   0.6060008
52             nem           tomeg  -0.5753438
```

Mátrixok

- Adat-mátrix
- Korreláció-mátrix
  - A változók közötti lineáris kapcsolatot mutatja
  - Standardizált kovariancia
  - Szimmetrikus
  - Négyzetes
  - A főátlóban 1-esek állnak (minden változó korrelál önmagával), a főátlóra tükrösen szimmetrikus

$$r = \frac{\mathrm{COV}_{xy}}{S_x S_y}$$

- Variancia-kovariancia mátrix Σ
  - Négyzetes és szimmetrkus
  - ◦
  - A főátlóban az egyes változók varianciái állnak, a többi elem a megfelelő változók kovarianciái

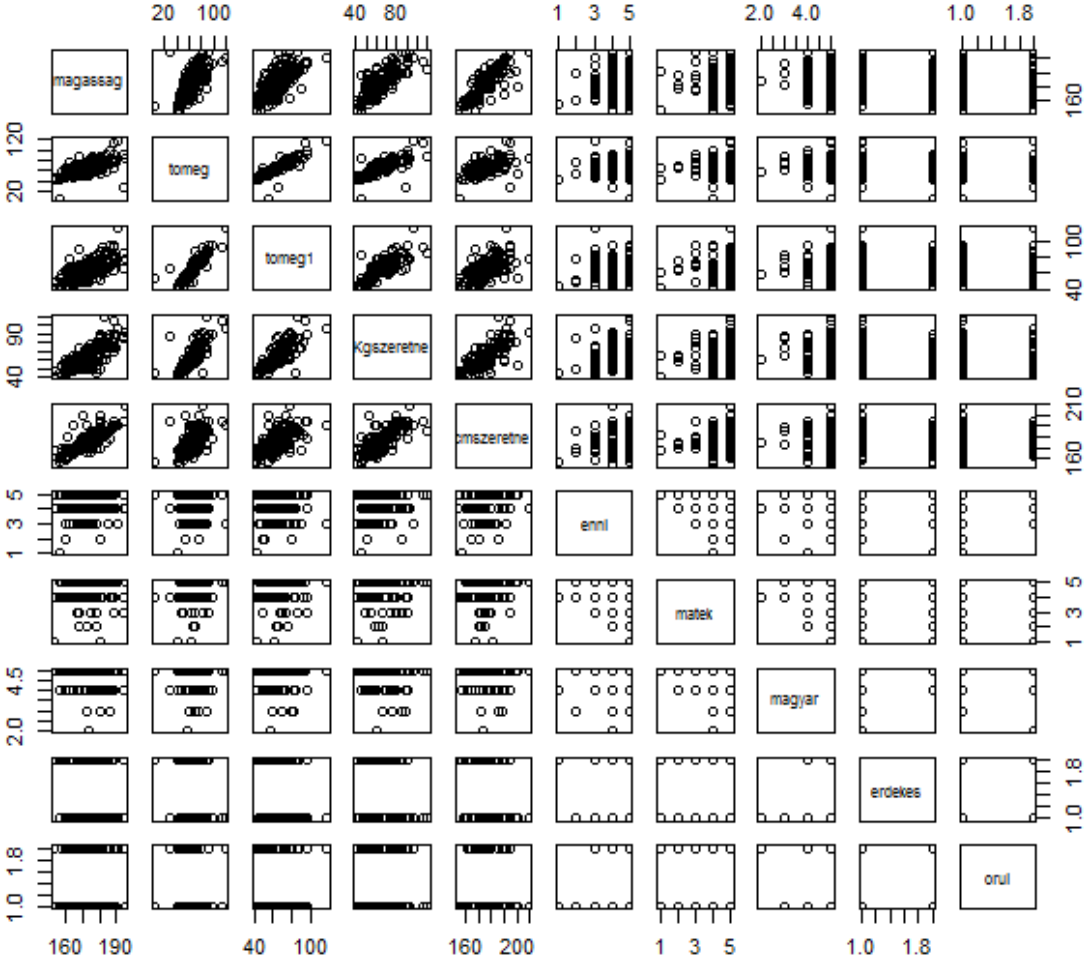
```
> cor(vars, use="complete.obs", method="pearson")   # pearson korrelációs mátrix
              magassag       tomeg      tomeg1  Kgszeretne  cmszeretne         enni        matek       magyar      erdekes         orul
magassag    1.00000000  0.66155465  0.68899707  0.79294131  0.83806844  0.177488585  0.03916975 -0.182724563 -0.124449713 -0.056668655
tomeg       0.66155465  1.00000000  0.89281012  0.80570019  0.60827264  0.067599876  0.07139018 -0.090914987 -0.116977516 -0.106856318
tomeg1      0.68899707  0.89281012  1.00000000  0.84333162  0.60600084  0.072084756  0.04639394 -0.103502785 -0.098225721 -0.082050382
Kgszeretne  0.79294131  0.80570019  0.84333162  1.00000000  0.76309418  0.156041572  0.06498027 -0.194171041 -0.179613082 -0.102635136
cmszeretne  0.83806844  0.60827264  0.60600084  0.76309418  1.00000000  0.213439098  0.03793651 -0.170037256 -0.193325852 -0.084542357
enni        0.17748859  0.06759988  0.07208476  0.15604157  0.21343910  1.000000000 -0.08145307  0.064796586  0.009574724  0.038039116
matek       0.03916975  0.07139018  0.04639394  0.06498027  0.03793651 -0.081453069  1.00000000  0.141342639 -0.046774069 -0.011991414
magyar     -0.18272456 -0.09091499 -0.10350279 -0.19417104 -0.17003726  0.064796586  0.14134264  1.000000000  0.099318985 -0.008201988
erdekes    -0.12444971 -0.11697752 -0.09822572 -0.17961308 -0.19332585  0.009574724 -0.04677407  0.099318985  1.000000000  0.389507035
orul       -0.05666865 -0.10685632 -0.08205038 -0.10263514 -0.08454236  0.038039116 -0.01199141 -0.008201988  0.389507035  1.000000000
> cov(vars, use="complete.obs")                    #kovariancia mátrix
             magassag       tomeg      tomeg1  Kgszeretne cmszeretne         enni        matek      magyar      erdekes         orul
magassag   76.5954520  76.7796051  72.0010152  87.4179737 66.8573423  1.216105781  0.247220953 -0.77046850 -0.538399066 -0.219785798
tomeg      76.7796051 175.8560479 141.3700827 134.5892340 73.5266992  0.701817167  0.682731841 -0.58085884 -0.766813867 -0.627963048
tomeg1     72.0010152 141.3700827 142.5737272 126.8458708 65.9570326  0.673849043  0.399497487 -0.59542663 -0.579767524 -0.434165778
Kgszeretne 87.4179737 134.5892340 126.8458708 158.6778844 87.6202477  1.538855896  0.590299985 -1.17841734 -1.118420385 -0.572940460
cmszeretne 66.8573423  73.5266992  65.9570326  87.6202477 83.0876098  1.523146033  0.249378204 -0.74673874 -0.871097914 -0.341505507
enni        1.2161058   0.7018172   0.6738490   1.5388559  1.5231460  0.612913050 -0.045987513  0.02444038  0.003705396  0.013197300
matek       0.2472210   0.6827318   0.3994975   0.5903000  0.2493782 -0.045987513  0.520075123  0.04910918 -0.016674280 -0.003832293
magyar     -0.7704685  -0.5808588  -0.5954266  -1.1784173 -0.7467387  0.024440384  0.049109182  0.23212020  0.023653622 -0.001751180
erdekes    -0.5383991  -0.7668139  -0.5797675  -1.1184204 -0.8710979  0.003705396 -0.016674280  0.02365362  0.244353079  0.085325618
orul       -0.2197858  -0.6279630  -0.4341658  -0.5729405 -0.3415055  0.013197300 -0.003832293 -0.00175118  0.085325618  0.196385970
```

Többváltozós adatok szemléltetése

- Mátrix plot a korrelációkról

```
a=read.csv("orvkerd2011.csv")
#############kitöröljük a hiányzó értékeket
        tartalmazó eseteket

ha=na.omit(a)
attach(ha)
#####mátrix plot#############
vars<-as.data.frame(cbind(magassag, tomeg, tomeg
        Kgszeretne, cmszeretne, enni,matek,magyar,
        erdekes, orul))
vars
plot(vars)
```



Többváltozós adatok szemléltetése

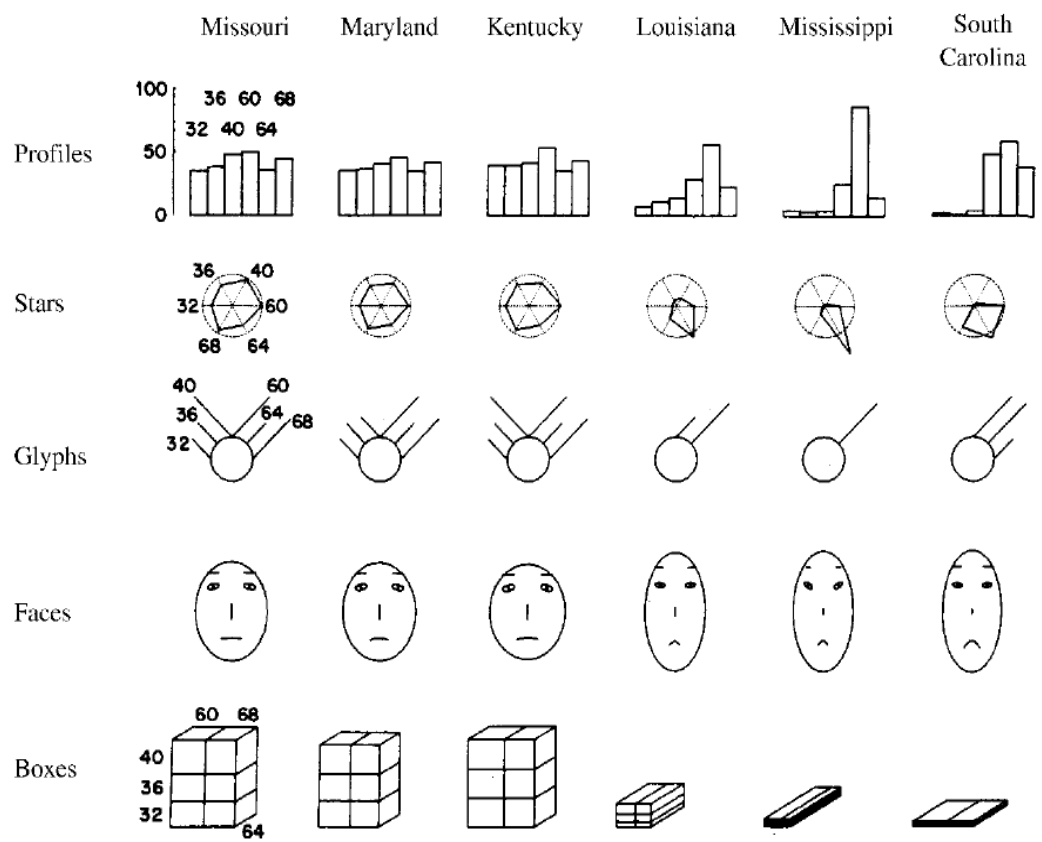
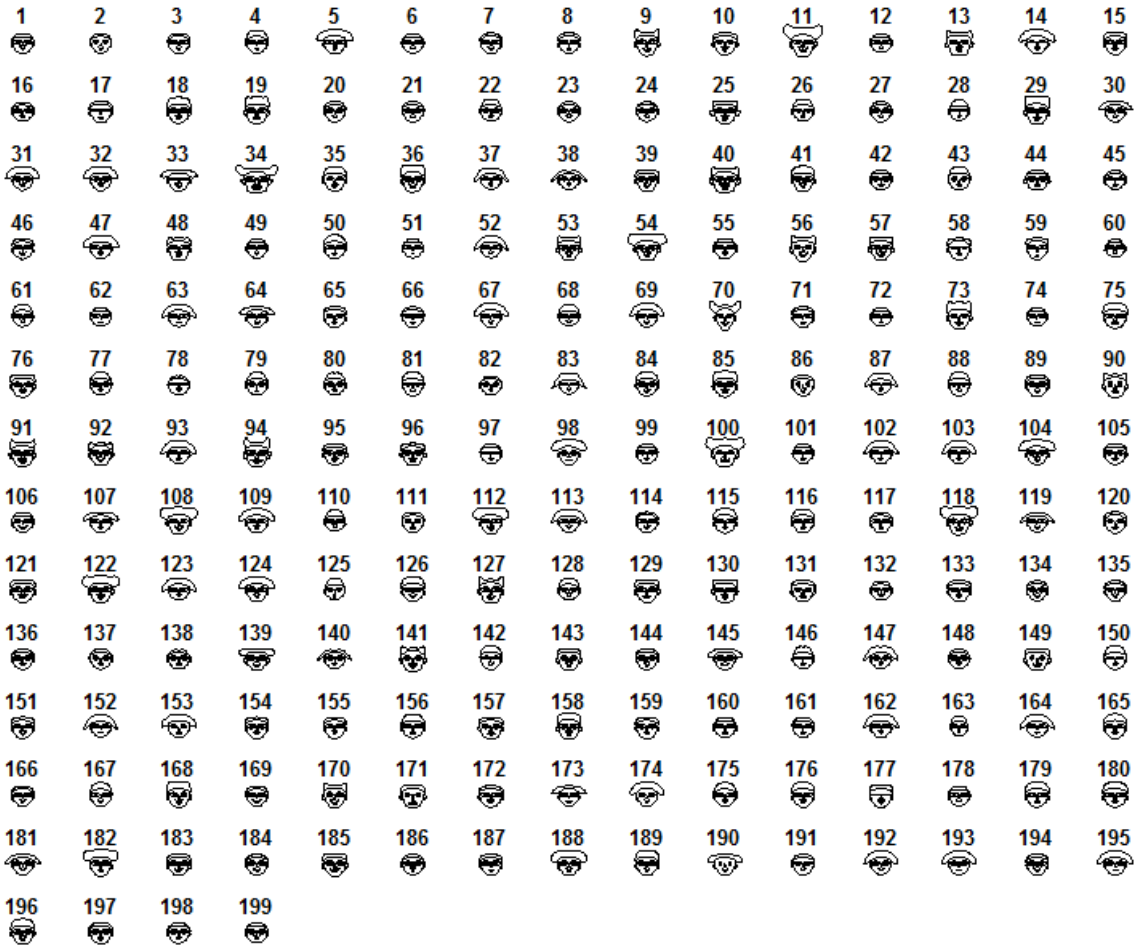


Figure 3.8. Profiles, stars, glyphs, faces, and boxes of percentage of Republican votes in six presidential elections in six southern states. The radius of the circles in the stars is 50%. Assignments of variables to facial features are 1932, shape of face; 1936, length of nose; 1940, curvature of mouth; 1960, width of mouth; 1964, slant of eyes; and 1968, length of eyebrows. (From the *Journal of the American Statistical Association*, 1981, p. 262.)

R

```
> stars(vars)
> library(aplpack)
Loading required package: tcltk
Loading Tcl/Tk interface ... done
Warning message:
package ‘aplpack’ was built under R v
> faces(vars,face.type=0)
effect of variables:
 modified item       Var
 "height of face   " "magassag"
 "width of face    " "tomeg"
 "structure of face" "tomeg1"
 "height of mouth  " "Kgszeretne"
 "width of mouth   " "cmszeretne"
 "smiling          " "enni"
 "height of eyes   " "matek"
 "width of eyes    " "magyar"
 "height of hair   " "erdekes"
 "width of hair    "  "orul"
 "style of hair    "  "magassag"
 "height of nose   "  "tomeg"
 "width of nose    "  "tomeg1"
 "width of ear     "  "Kgszeretne"
 "height of ear    "  "cmszeretne"
```

Statistica for Windows



Explanation of parameters: 1-height of face, 2-width of face, 3-shape of face, 4-height of mouth, 5-width of mouth, 6-curve of smile, 7-height of eyes, 8-width of eyes, 9-height of hair, 10-width of hair, 11-styling of hair, 12-height of nose, 13-width of nose, 14-width of ears, 15-height of ears.
For painting elements of a face the colors of are found by averaging of sets of variables: (7,8)-eyes:iris, (1,2,3)-lips, (14,15)-ears, (12,13)-nose, (9,10,11)-hair, (1,2)-face.

Elemszám

- **A többváltozós módszerek nagyon "adatigényesek". Az elemszám:**
  - **a változók számának 10-szerese**
  - **a változók számának 3-szorosa**
  - **egyre terjednek a minden egyes módszerhez a módszer adottságait figyelembe vevő elemszám-becslések**

Multikollinearitás

- Ha a független változók korrelálnak, problémák lehetnek a regressziós együtthatók becslésekor, azaz minél jobban korrelálnak, annál nagyobbak lesznek a regressziós együtthatók standard hibái.
- A modell struktúrájának akár enyhe változtatásánál megváltozhatnak a paraméterek előjelei vagy nagyságuk.

Csak néhány eljárást tanulunk

- Csoportok „szétválasztása” a csoportok ismeretében, osztályozás, szignifikanciavizsgálat:
  - diszkriminancia analízis (MANOVA)
  - logisztikus regresszió (külön előadás)
- Dimenziócsökkentés (lényegkiemelés):
  - Faktoranalízis
- Csoportok felfedezése, megtalálása az adatok között változóik alapján vagy a változók között korrelációik alapján:
  - Cluster analízis
- ROC analízis
  - Osztályozási problémák, diagnosztikus tesztek esetén a modell diagnosztizáló képességét írja le, illetve segít az „elvágópont” megtalálásában.

A változók számának csökkentése

- **Stepwise algoritmussal**
- **előkészítő tevékenység lehet egy többváltozós analízis során**
  - **faktoranalízis**
  - **cluster analízis**

Faktoranalízis

- több, egymással korreláló változóból a korrelációs mátrix alapján új, közvetlenül nem megfigyelhető, hipotetikus konstrukciókat, ún. faktorokat hoz létre. Ezeknek a faktoroknak a száma szerencsés esetben sokkal kevesebb, mint az eredeti változóké, és viszonylag kis információveszteséggel jól leírják az eredeti adathalmazt.
- Az eredeti változók és a faktorok közötti kapcsolatot az un. faktorsúlyokkal mérjük. Ezek lényegében a változók és a faktorok közötti korrelációs együtthatók. Minél közelebb van az abszolút értékük 1-hez, annál szorosabb a kapcsolat az adott faktorral. Általában mátrix formában szokták megadni, melynek sorai a változók, oszlopai a faktorok. Több módszer ismeretes a faktorok extrakciójára (főkomponens analízis, maximum likelihood módszer, stb.)

	X_1	X_2	...	X_j	...	X_p
1.egyed						
2.egyed						
...						
i.egyed				x_{ij}		
...						
n. egyed						

F_1 F_2 F_3

Példa

Egészséges tanulólányok kérdőíves adatainak elemzése

(Dr. Pethő Zoltán, Neuropszichiátriai Rehabilitációs Osztály)

A kérdőíves szűrővizsgálat Debrecenben és Szegeden az Egészségügyi Szakközépiskola I-IV. osztályaiban

Általános demográfiai, szociális és kulturális adatok

NEM Neme

LAKIK Hol lakik
- 1 otthon
- 2 kollégiumban
- 3 albérletben
- 4 egyéb válasz

LAK Hol élt 14 éves koráig
- 1 Budapesten
- 2 megyeszékhely
- 3 albérletben
- 4 falu
- 5 tanya
- 6 egyéb

APIS Apja legmagasabb iskolai végzettsége
- 1 8 ált.-nal kevesebb
- 2 8 ált.
- 3 szakmunkásképző
- 4 szakközépiskolai érettségi
- 5 gimnáziumi érettségi
- 6 felsőfokú végzettség

ANYIS Anyja legmagasabb isk. v.

TESTV testvérek száma

SZOP hány hónapos koráig szopott

SZOM Volt-e levert időszaka életében
- 1 igen
- 2 nem

ONGY Öngyilkossági késztetés
- 1 sohasem volt
- 2 foglalkoztam
- 3 volt kísérletem

ALK Az alkoholfogyasztásról
- 1 nem iszom alkoholt
- 2 társaságban iszom
- 3 rendszeresen, keveset iszom
- 4 egyszer lerészegedem
- 5 többször lerészegedem
- 9 nincs válasz

CIGI Cigaretták száma/nap

CIGEV Hány éve dohányzik

KAVE Hányszor iszik kávét

HAVIE Mikor volt az elsô havivérzése

HAVIR Hány éves kora óta rendszeres a havivérzése

JSULY Jelenlegi testsúlya /kg/

KSULY Kivánatosnak tartott testsúlya /kg/

MAGAS Jelenlegi testmagassaga /cm/

MISULY Minimális testsúlya /kg/

MAXSULY Maximális testsúlya /kg/

JOSULY Mi a véleménye a testsúlyáról

GYORSV Volt-e 5 kg-nál nagyobb testsúlyváltozása

KEZEL Kezelték-e étvágyzavar miatt

AGGOD Aggódtak-e mások soványsága vagy kövérsége miatt

ENNI Mennyire szeret enni

Forgatás után elég egyértelműek a faktorok, (lehetőleg minden változó egy és csak egy faktorhoz tartozzon), el is lehet őket nevezni

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,617	20,134	20,134	2,173	16,716	16,716
2	1,690	13,000	33,134	1,775	13,653	30,369
3	1,482	11,402	44,536	1,632	12,551	42,920
4	1,251	9,624	54,160	1,278	9,830	52,751
5	1,042	8,014	62,174	1,225	9,423	62,174
6						
7						
8						
9						
10						
11						
12						
13						

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
magasság Testmagassága	,857	,207			
jsuly Jelenlegi testsúlya	,845				
nem	-,791	-,179		-,187	
apis Az apa iskolai végzettsége	,123	,888			
anyis Az anya iskolai végzettsége	,117	,884			
pr1 Próbálkozott-e már droggal			,840		
dr1 Találkozott-e már droggal		,120	,833		
depr Volt már valaha depressziós?		,238	-,431		-,249
ksuly Kívánatosnak tartott testsúlya	,213			,669	-,236
enni Mennyire szeret enni?				-,659	
kor		-,185	,195	,352	,341
ongyilk Öngyilkossági késztetés				-,177	,807
alkohol				,447	,569

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

F1: testméretek fakt

F2: szülők faktora

F3: drog

F4: evés és testsúly

További lehetőségek

- Ha a faktorok által magyarázott összvariancia aránya elég nagy és a faktorok jól értelmezhetők (ez nincs mindig így), akkor a keletkezett faktorok együtthatóit megbecsülve új változóként lehet őket elmenteni és a továbbiakban ezekkel dolgozni;
- Vagy a ha az eredeti változókkal szeretnénk tovább dolgozni, minden faktorból kiválaszthatjuk az őt legjobban képviselő eredeti változót (lehetőleg a legnagyobb súlyút)

Cluster analízis

- A cluster analízis megfigyelési egységek közötti kapcsolatot vagy pedig változók közötti kapcsolatot vizsgál, azt nézi, hogy a megfigyelt adatokat hogyan lehet struktúrákba csoportosítani, rendszerezni. Egy ilyen rendszerezési lehetőség a biológiában ismert rendszertan. Eszerint az ember főemlős, emlős, gerinces, stb. Minél messzebb megyünk az összevonás szintjében, annál kevésbé hasonlók az egy osztályba tartozó tagok.
- Meg kell jegyezni, hogy a cluster analízis nem jelent semmilyen statisztikai szignifikanica vizsgálatot. Lényegében a cluster analízis különböző algoritmusok “gyűjteménye”, és így nem tipikus statisztikai próba. A szokásos statisztikai próbáktól eltérően, a cluster analízis módszereket akkor alkalmazzuk legtöbbször, ha nincs semmi a priori hipotézisünk, a kutatás kezdeti fázisában vagyunk. Így a statisztikai szignifikancia-vizsgálat a tradicionális értelemben nem megfelelő itt, még akkor sem, ha esetleg p-értékek szerepelnek az eredmények között (K-közép eljárás) .

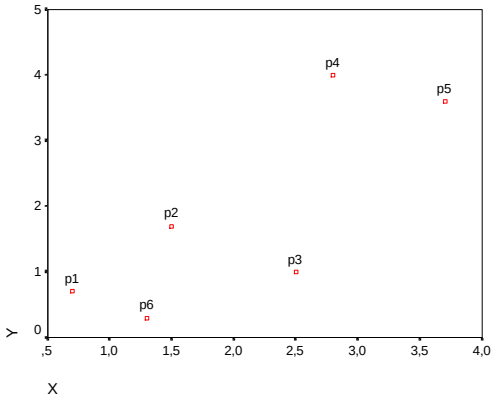
- A cluster eljárások lehetnek **hierarchikus vagy dinamikus** eljárások,
- más szempontból clusterezhetünk **megfigyelési egységeket (objektumokat) vagy változókat**.

Hierarchikus módszer

- **1. Megfigyelési egységek clusterezése**:
- Adott több változóval jellemzett megfigyelési egységeknek egy halmaza (olyan statisztikai minta, ahol minden egyedről többféle adat áll rendelkezésre). A cél a megfigyelései egységek (egyedek, objektumok) osztályokba sorolása a közöttük levő "kapcsolat" alapján. A kapcsolat jelenthet hasonlóságot vagy távolságot. Az egységek között valamilyen módon "távolságot" számítunk. Ez lehet Euklideszi távolság, vagy abszolút eltérés, stb. .E távolságokat egy mátrixba rendezzük. Adott tehát minden objektumnak mindegyiktől mért távolsága. A cluster analízis ezek alapján rendezi csoportokba az egységeket .
- **A hierarchikus módszerek lépései**:
- E módszereknél kezdetben minden objektum egy önálló osztály (cluster). 1. lépés: két esetet összevonunk, ez egy új cluster lesz, a clusterek száma így eggyel csökken.
- lépés: vagy egy harmadik objektumot sorolunk a már meglevő clusterbe, vagy másik két objektumot vonunk össze egy új clusterbe. A clusterek száma megint 1-gyel csökken.
- Minden egyes lépésben tehát vagy egy eddig nem osztályozott objektumot vonunk be valamelyik létező clusterbe, vagy két létező clustert összevonunk. A már meglevő clustereket nem lehet szétvágni, így mindig nagyobb elemszámú clusterek keletkeznek minden egyes lépésnél, amíg a clusterek száma 1-gyel csökken.
- Az eljárást addig folytatjuk, míg végül egy cluster lesz.
- A folyamatot szokták un. dendogramon ábrázolni. Ez egy fához hasonló ábra, mely az összevonás lépéseit mutatja. A kutató dönthet, hogy az n cluster közül hányat óhajt megtartani

A pontok koordinátái:

PONT	X	Y
p1	,70	,70
p2	1,50	1,70
p3	2,50	1,00
p4	2,80	4,00
p5	3,70	3,60
p6	1,30	,30



Távolságmátrix:

Proximity Matrix

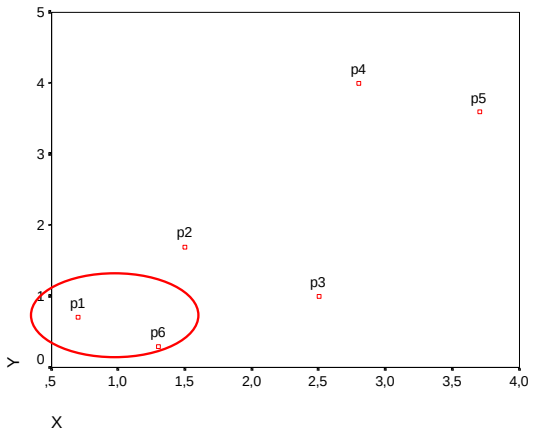
	Euclidean Distance					
Case	1:p1	2:p2	3:p3	4:p4	5:p5	6:p6
1:p1		1,281	1,825	3,912	4,173	,721
2:p2	1,281		1,221	2,642	2,907	1,414
3:p3	1,825	1,221		3,015	2,864	1,389
4:p4	3,912	2,642	3,015		,985	3,992
5:p5	4,173	2,907	2,864	,985		4,080
6:p6	,721	1,414	1,389	3,992	4,080	

This is a dissimilarity matrix

Első lépés: 1-6.

Proximity Matrix

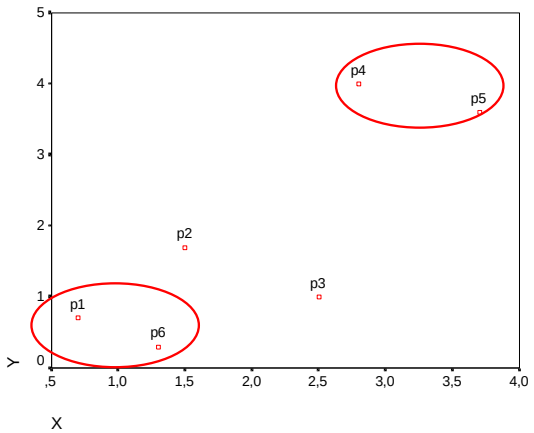
	Euclidean Distance					
Case	1-6:p1	2:p2	3:p3	4:p4	5:p5	
1-6:p1		1,281	1,389	3,912	4,080	
2:p2	1,281		1,221	2,642	2,907	
3:p3	1,389	1,221		3,015	2,864	
4:p4	3,912	2,642	3,015		,985	
5:p5	4,080	2,907	2,864	,985		



Második lépés: 4-5.

Proximity Matrix

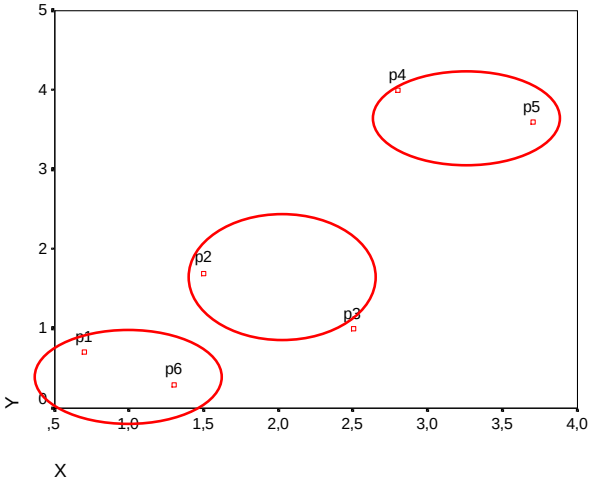
	Euclidean Distance				
Case	1-6:p1	2:p2	3:p3	4-5:p4	
1-6:p1		1,281	1,389	3,912	
2:p2	1,281		1,221	2,642	
3:p3	1,389	1,221		2,864	
4-5:p4	3,912	2,642	2,864		



Harmadik lépés: 2-3.

Proximity Matrix

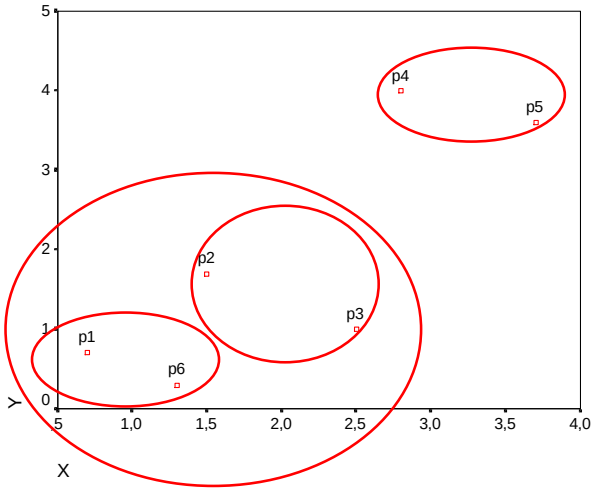
	Euclidean Distance			
Case	1-6:p1	2-3:p2	4-5:p4	
1-6:p1		1,281	3,912	
2-3:p2	1,281		2,642	
4-5:p4	3,912	2,642		



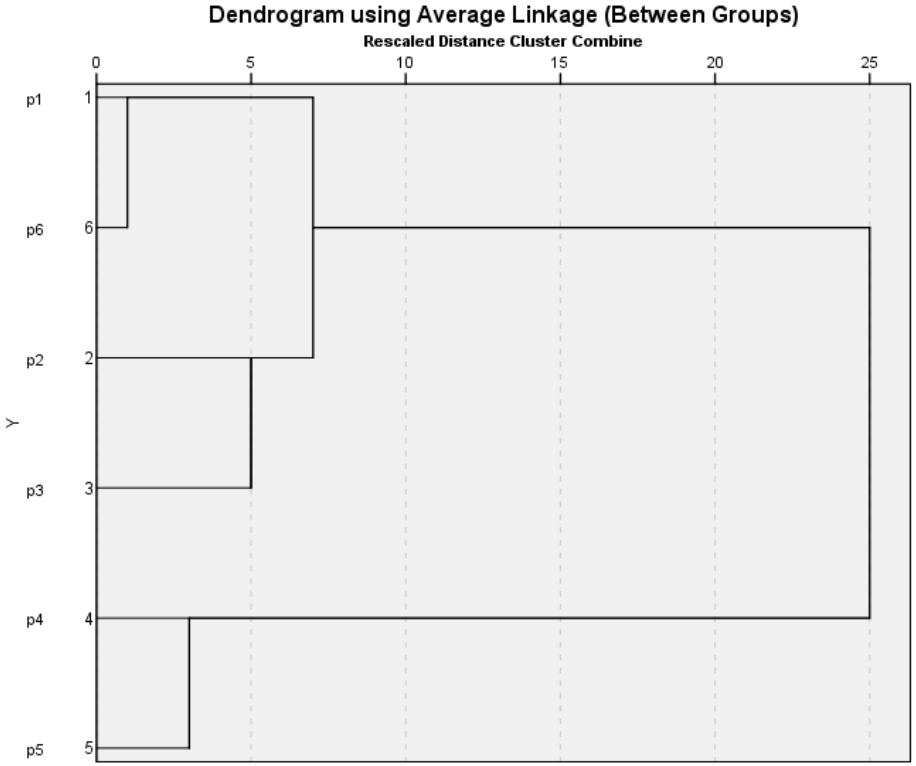
Negyedik lépés: 1-6 - 2-3.

Proximity Matrix

	Euclidean Distance		
Case	1-6-2-3:p1	4-5:p4	
1-6-2-3:p1		2,642	
4-5:p4	2,642		



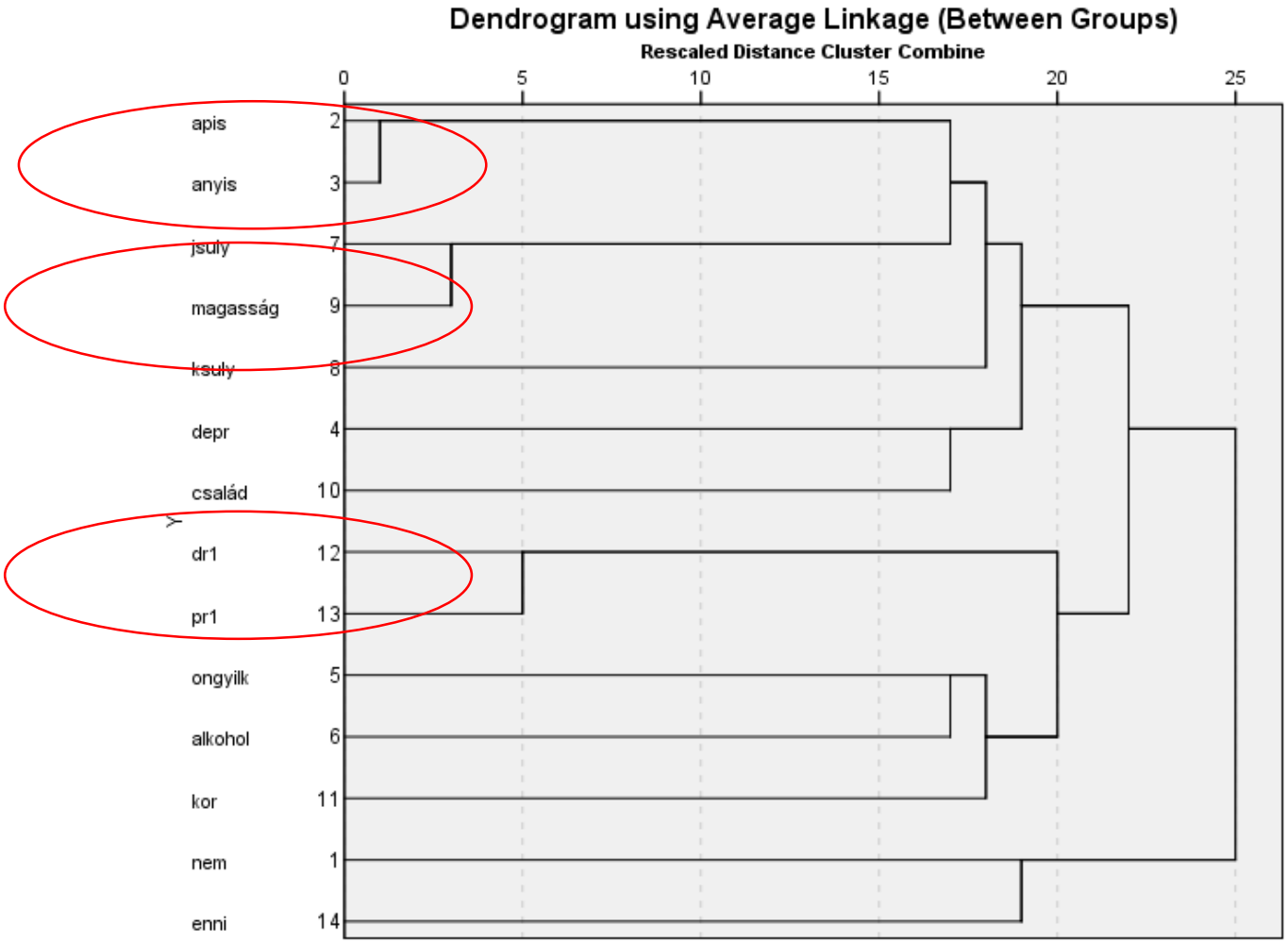
Az összevonás folyamata az ún. dendogramon szemléltethető.



A <u>változók</u> hierarchikus clusteranalízise a korrelációk alapján

- A változók közötti korrelációk azok „hasonlóságát” fejezik ki. A korrelációs mátrix alapján a változókat is clusterezhetjük.

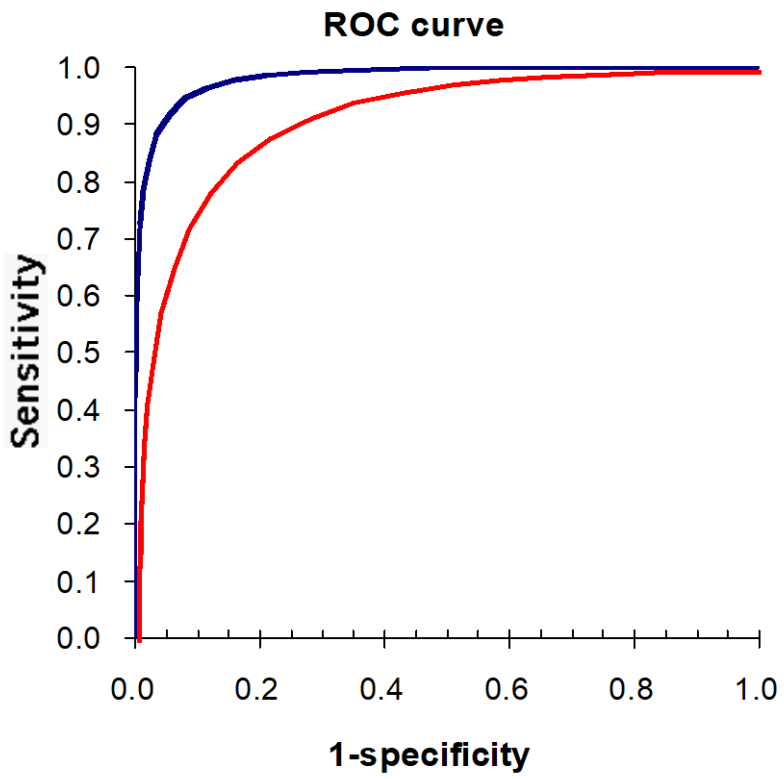
Példa a változók hierarchikus clusterérezésére a korrelációk alapján



A faktoranalízissel kapott faktorokhoz hasonló csoportokat kaptunk

A ROC-görbe
Hatásfokmérő karakterisztika

- ROC - Receiver operating characteristics curve
- Ha különböző mérési értékeknél (vagy folytonos változóknál küszöb-értékeknél) kiszámítjuk a szenzitivitást és a specifikusságot, majd a szenzitivitást az (1-specifikusság) függvényében ábrázoljuk, a pontokat összekötő vonal a hatásfokmérő karakterisztika (ROC görbe).



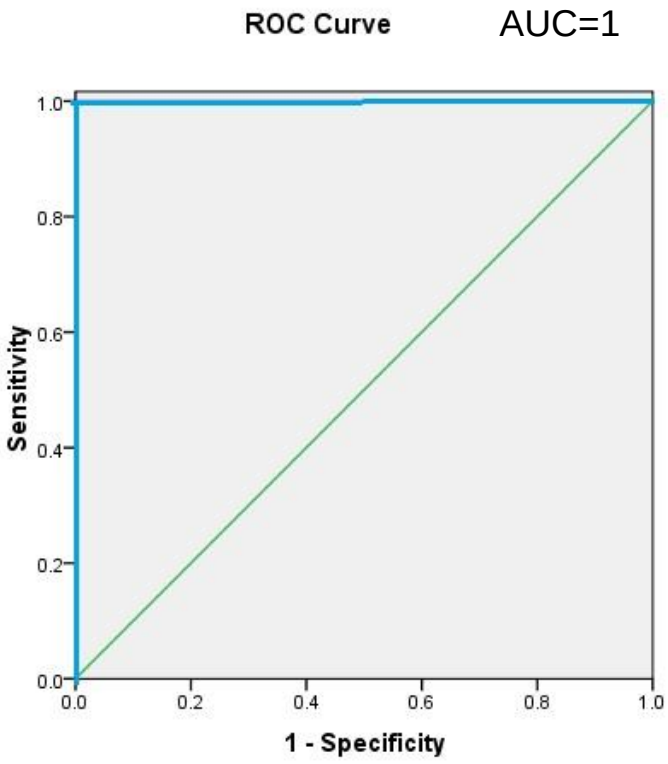
Az ábrán a kék görbével jellemzett teszt diagnosztizáló képessége jobb, mint a pirossal jelzett.

A ROC görbe tulajdonságai

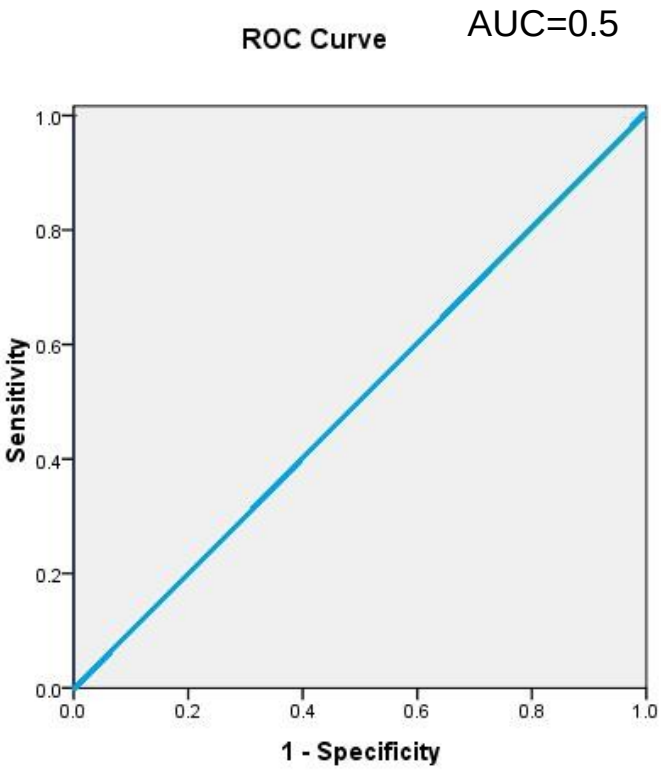
- Különböző küszöb-értékeknél mutatja a szenzitivitást és a specificitást
- Minél jobb a diagnosztikus teszt, annál jobban közelít a ROC görbe a bal-felső sarokhoz
- Tökéletes szétválasztás esetén a görbe két egymásra merőleges vonal, teljes egyezés esetén a görbe az átlóval egyenlő
- A görbe alatti terület a teszt diagnosztizáló képességének mérőszáma
  - ROC = 0,5 Nem használható teszt
  - ROC < 0,7 Gyenge szétválaszthatóság
  - 0,7 ≤ ROC < 0,8 Elfogadható a teszt
  - 0,8 ≤ ROC < 0,9 Jó diagnosztikus teszt
  - ROC ≥ 0,9 Kiváló diagnosztikus teszt

Két extremális ROC-görbe

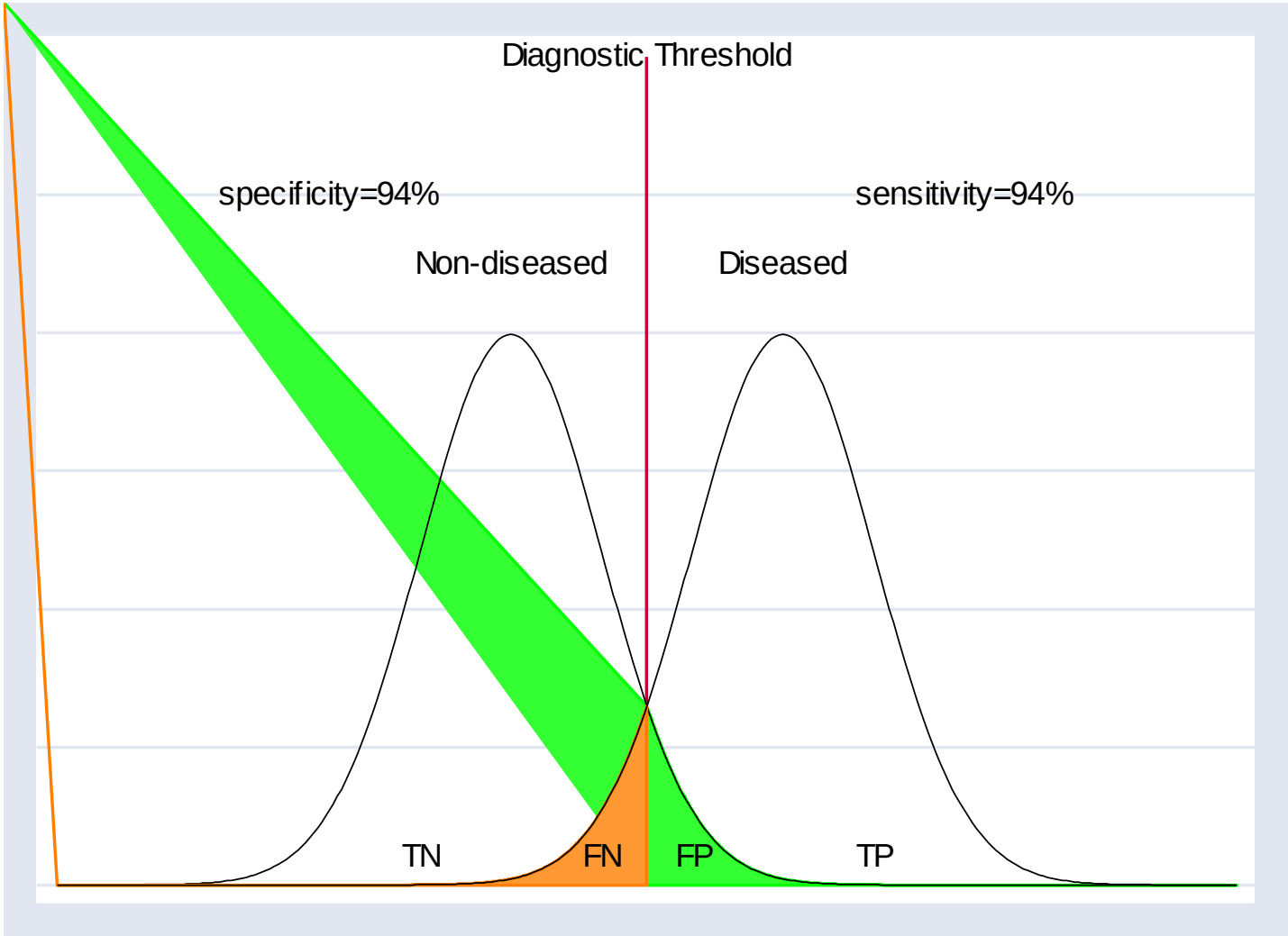
- Ideális osztályozás



- A legrosszabb osztályozás

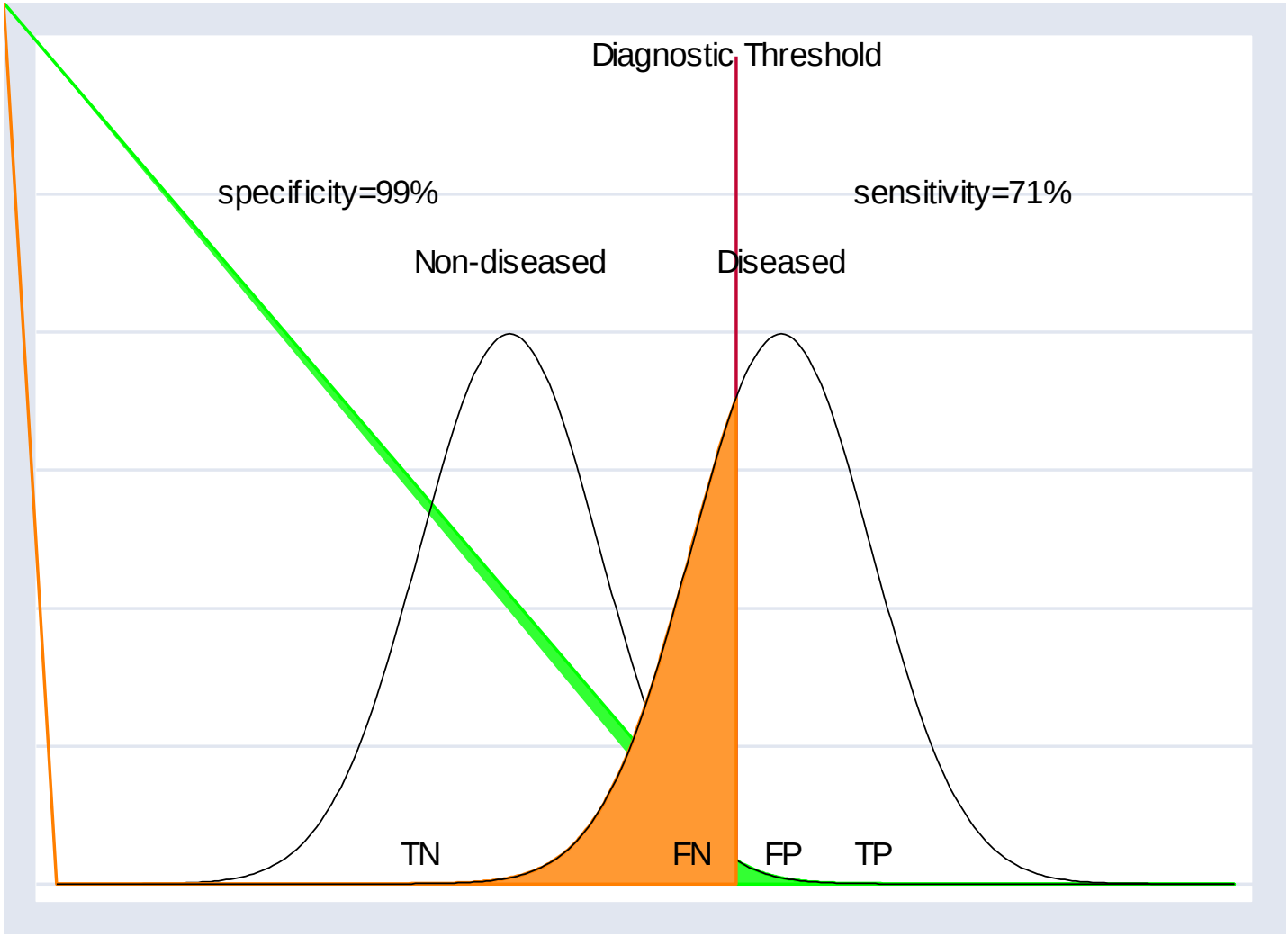


Folytonos változóval mért diagnosztikus teszteredmények eloszlása

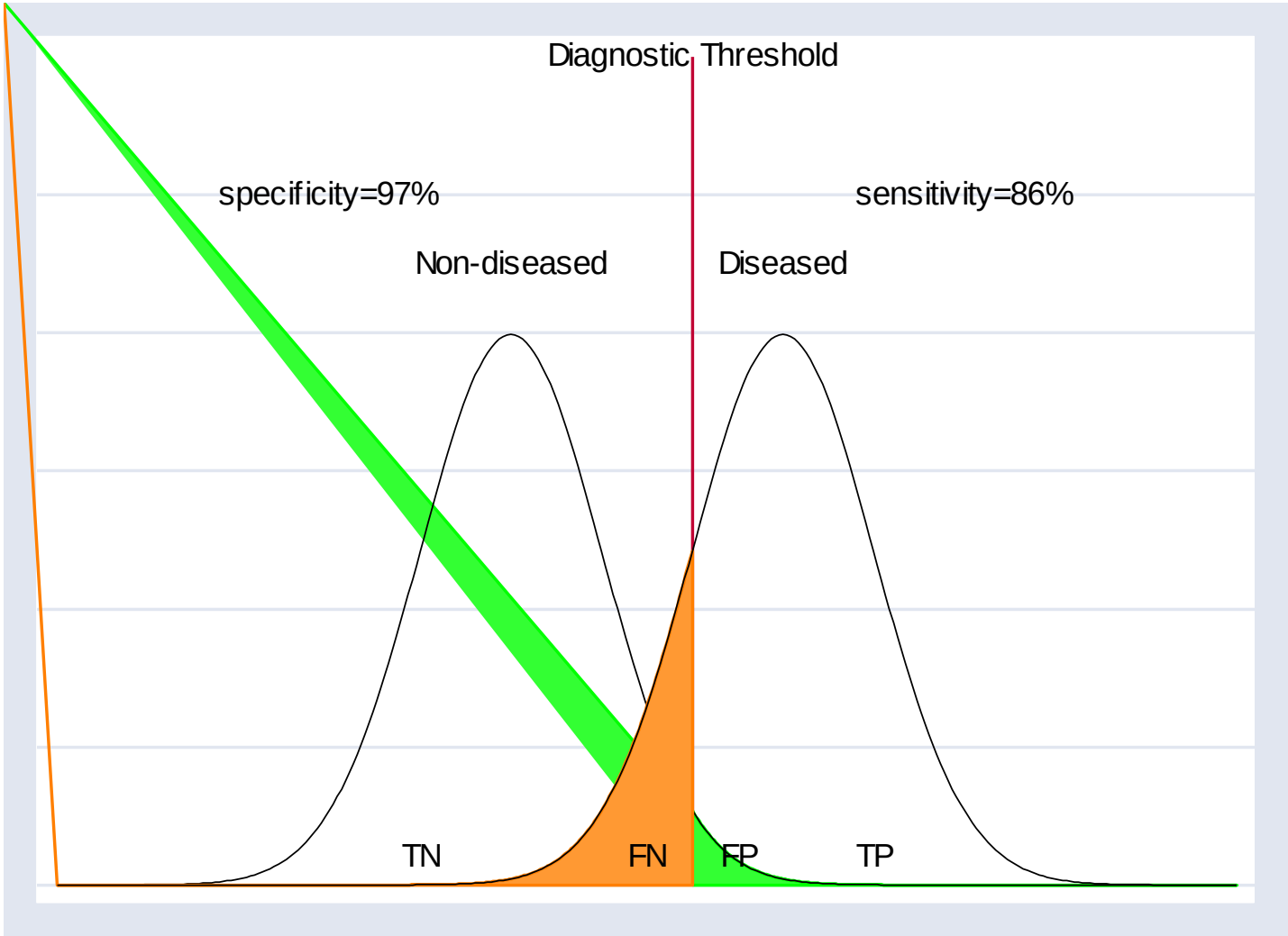


A küszöb-érték változtatásával változik a szenzitivitás és a specificitás

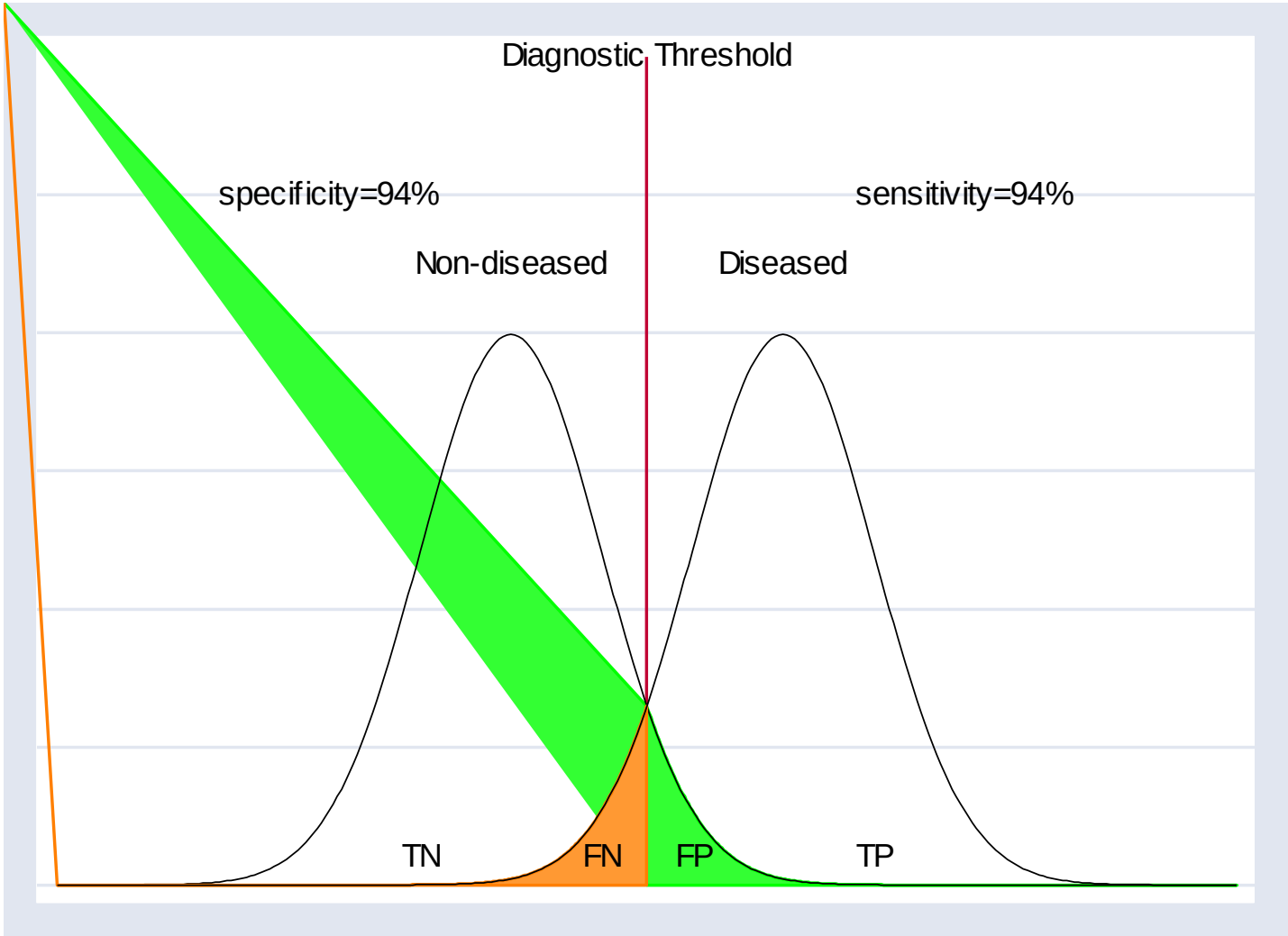
A küszöb-érték megválasztása



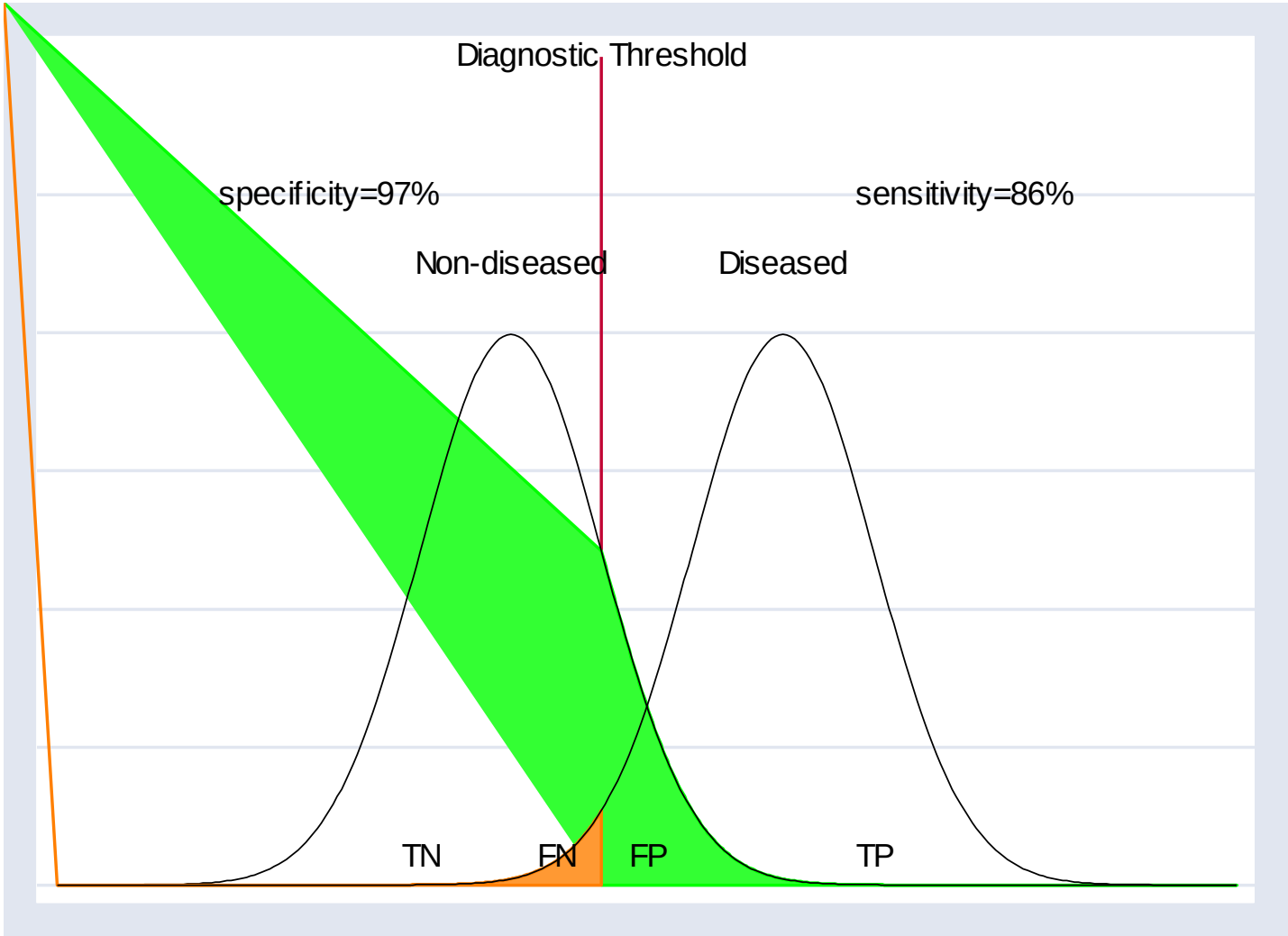
A küszöb-érték megválasztása



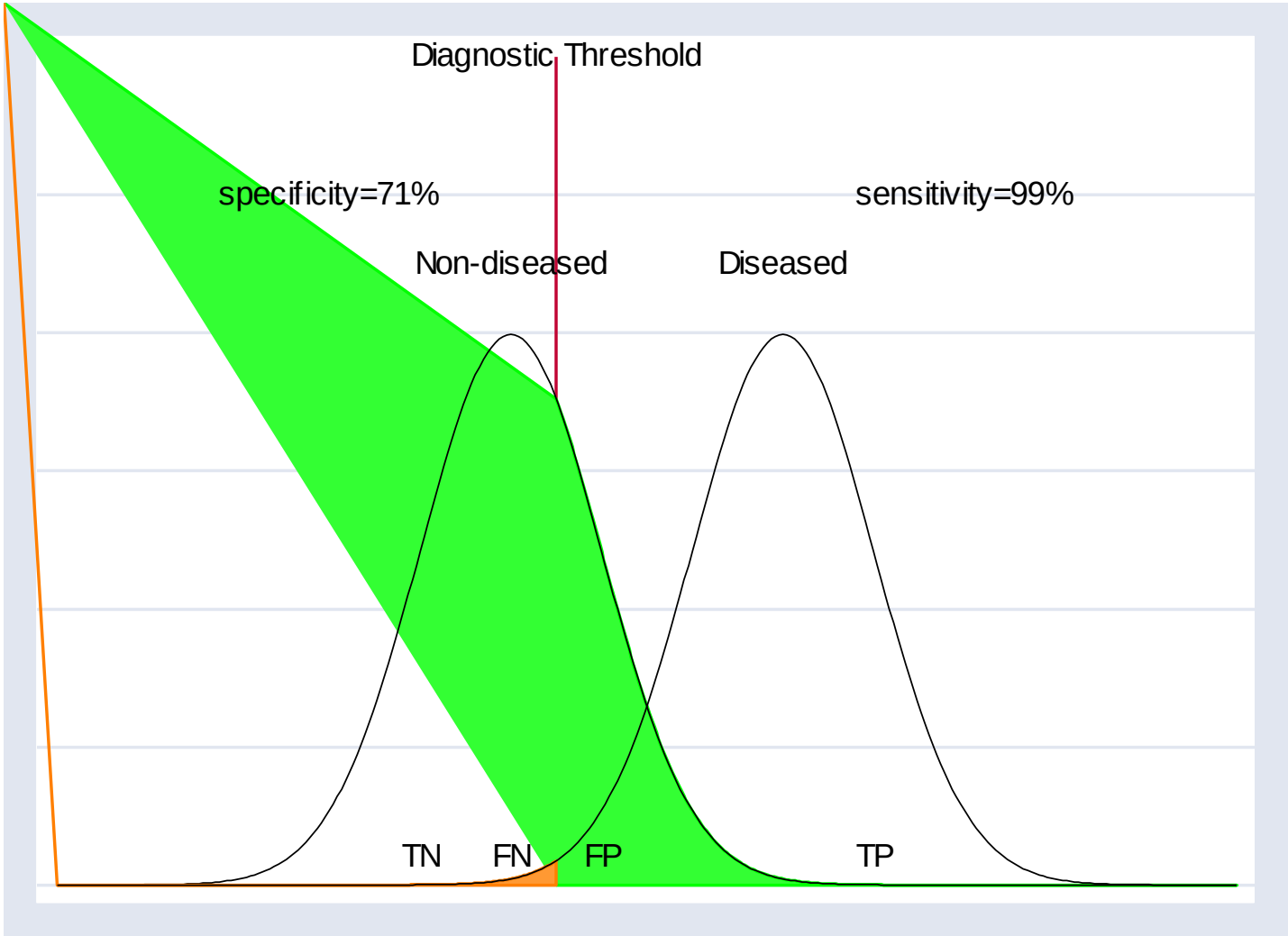
A küszöb-érték megválasztása



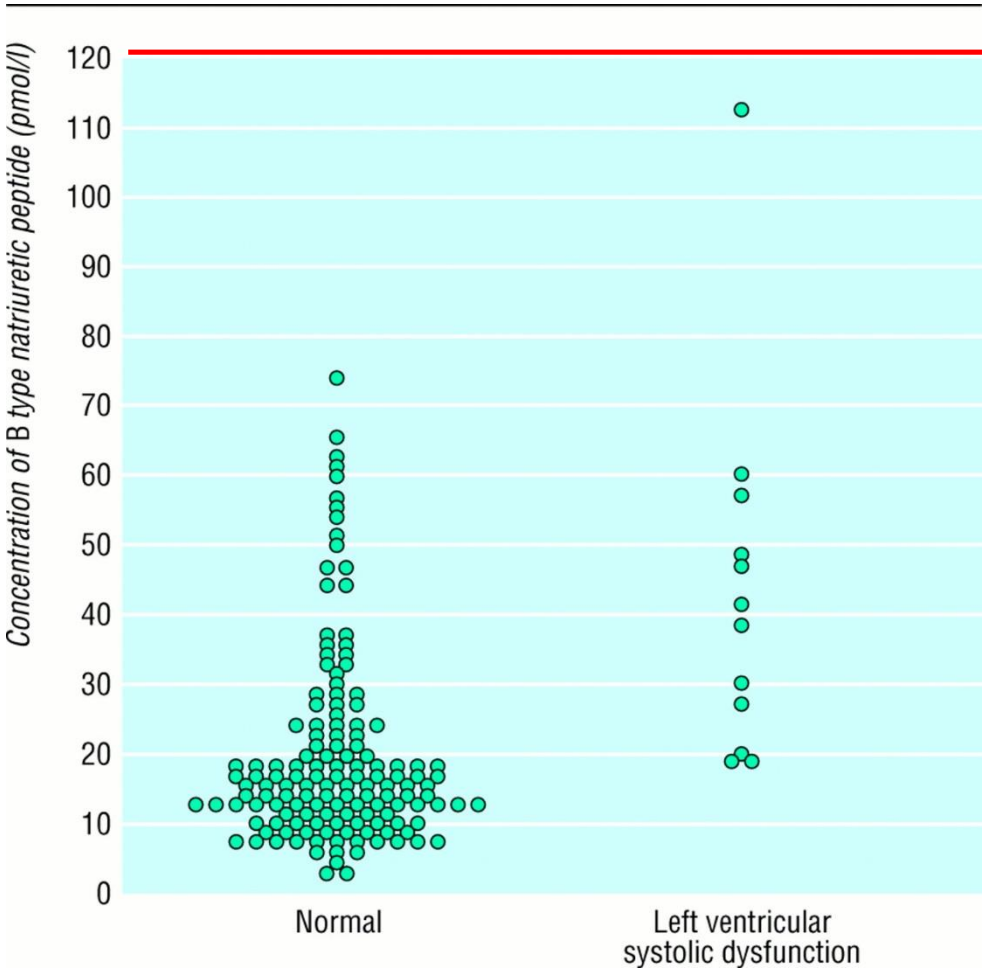
A küszöb-érték megválasztása



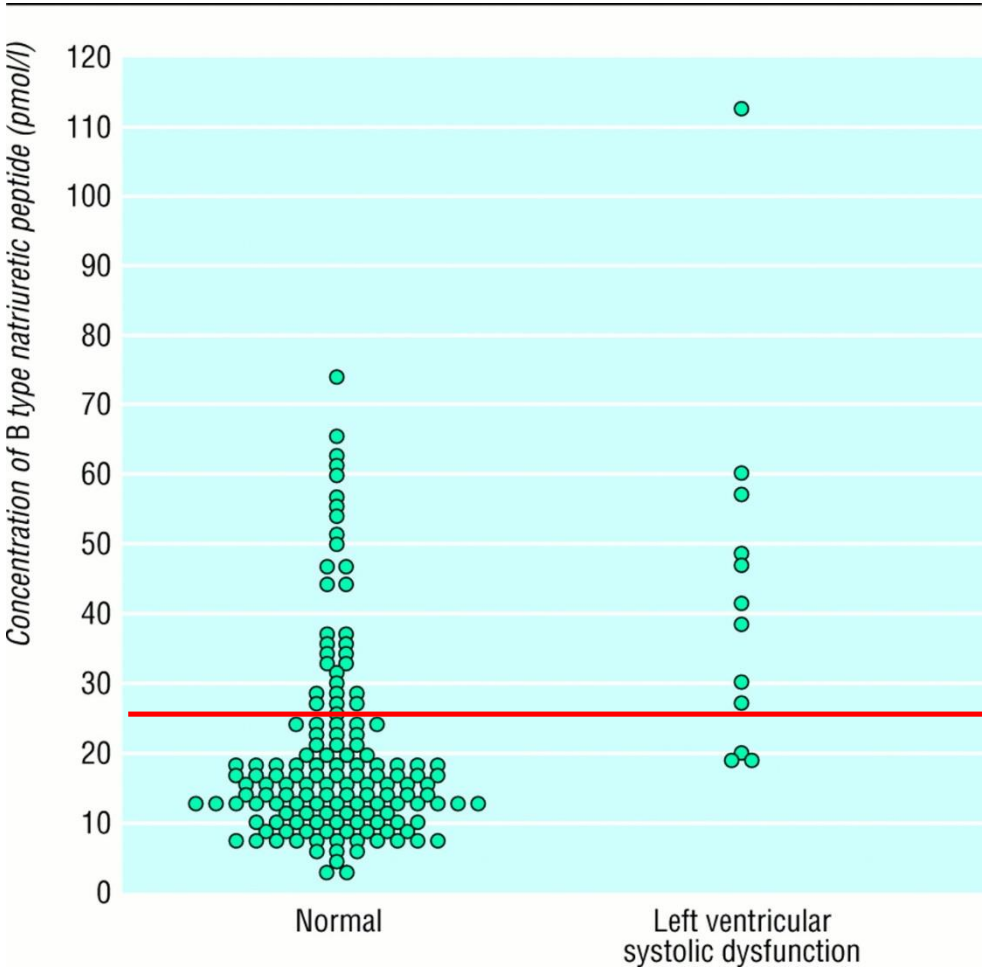
A küszöb-érték megválasztása



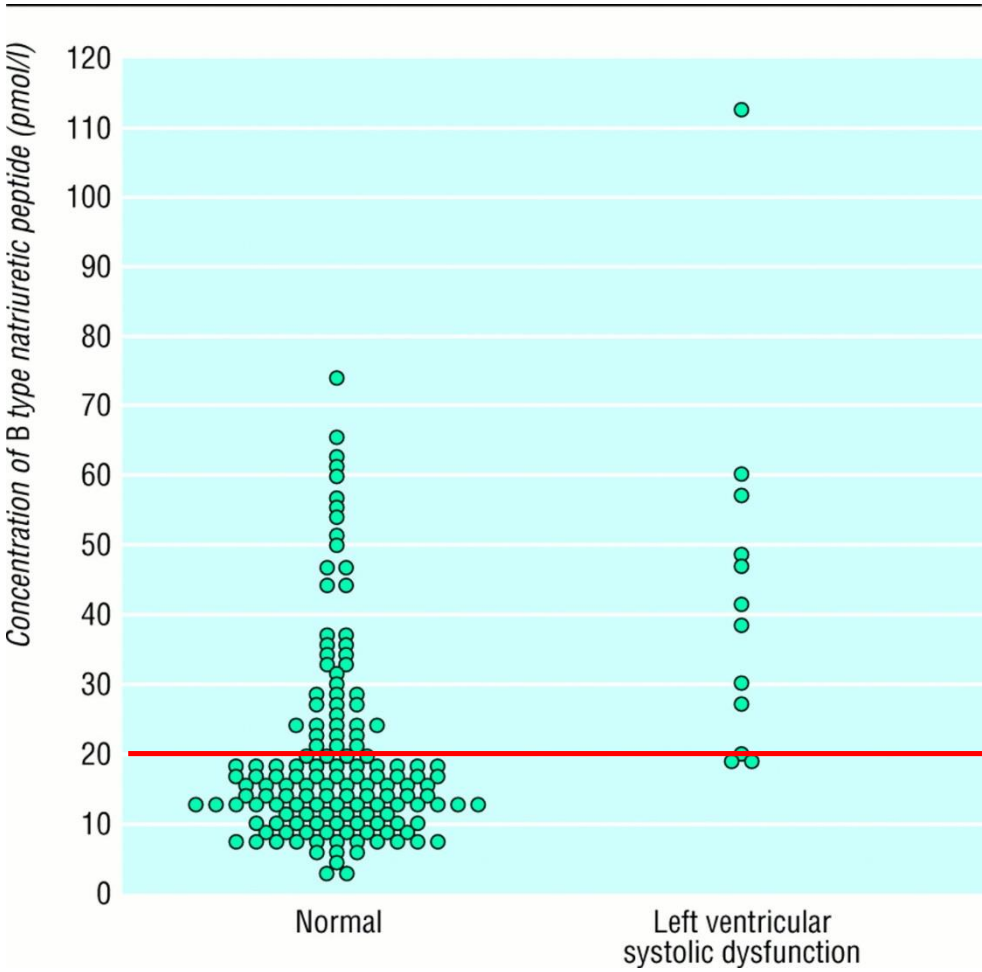
Küszöb	Szenz.	Spec.
≥ 9999	0%	100%
≥ 26.7	75%	80%
≥ 19.8	83%	70%
≥ 18.7	92%	65%
≥ 0	100%	0%



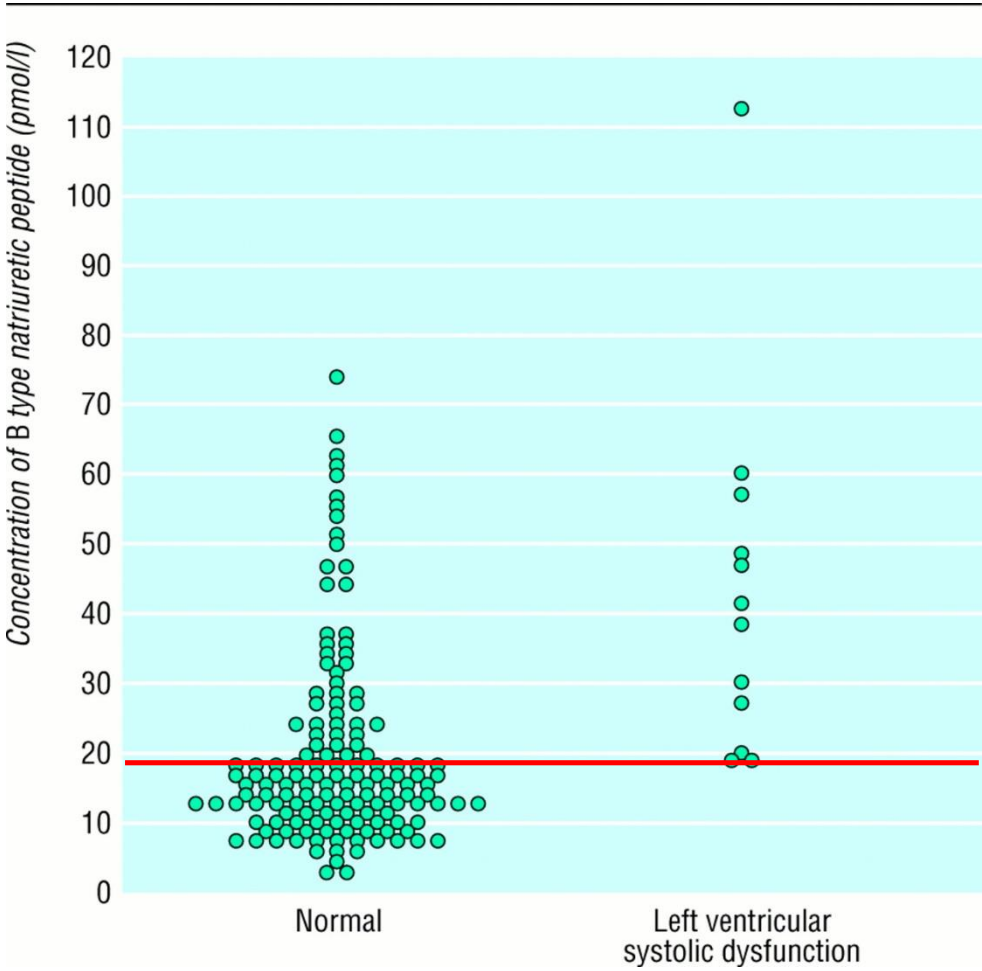
Küszöb	Szenz.	Spec.
≥ 9999	0%	100%
≥ 26.7	75%	80%
≥ 19.8	83%	70%
≥ 18.7	92%	65%
≥ 0	100%	0%



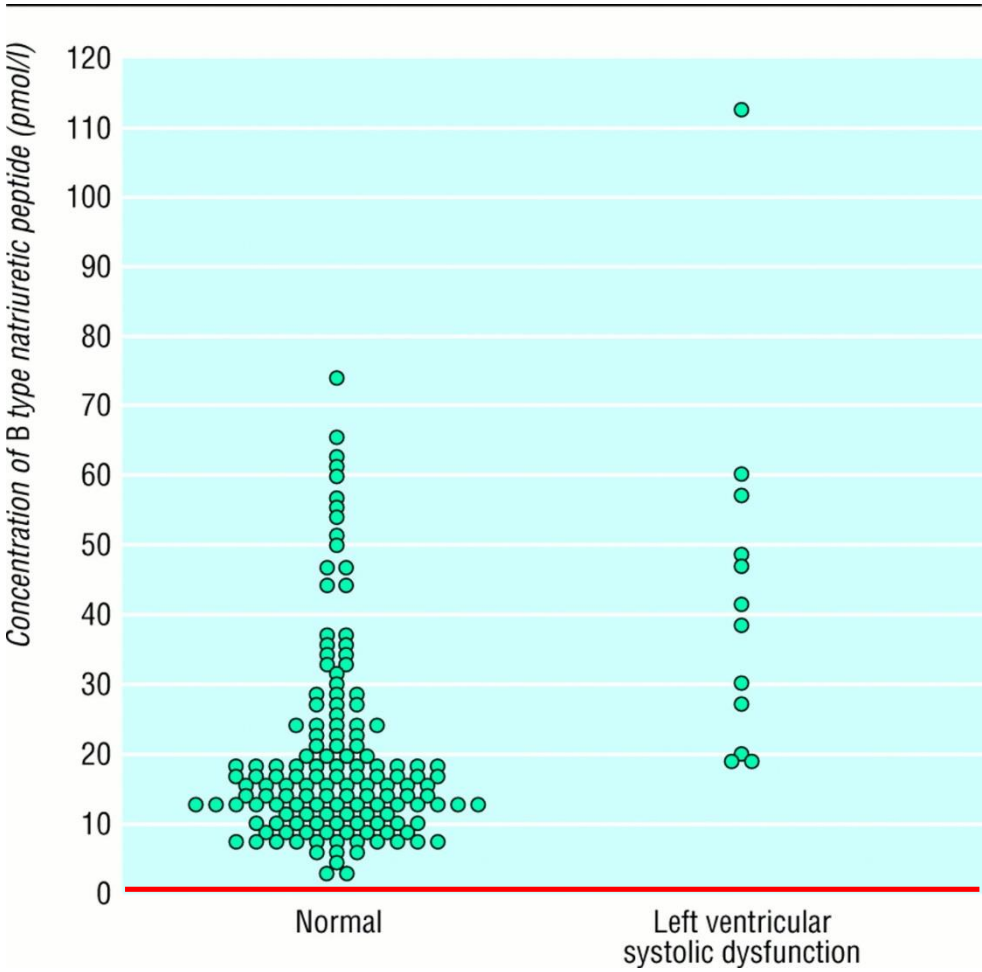
Küszöb	Szenz.	Spec.
≥ 9999	0%	100%
≥ 26.7	75%	80%
≥ 19.8	83%	70%
≥ 18.7	92%	65%
≥ 0	100%	0%



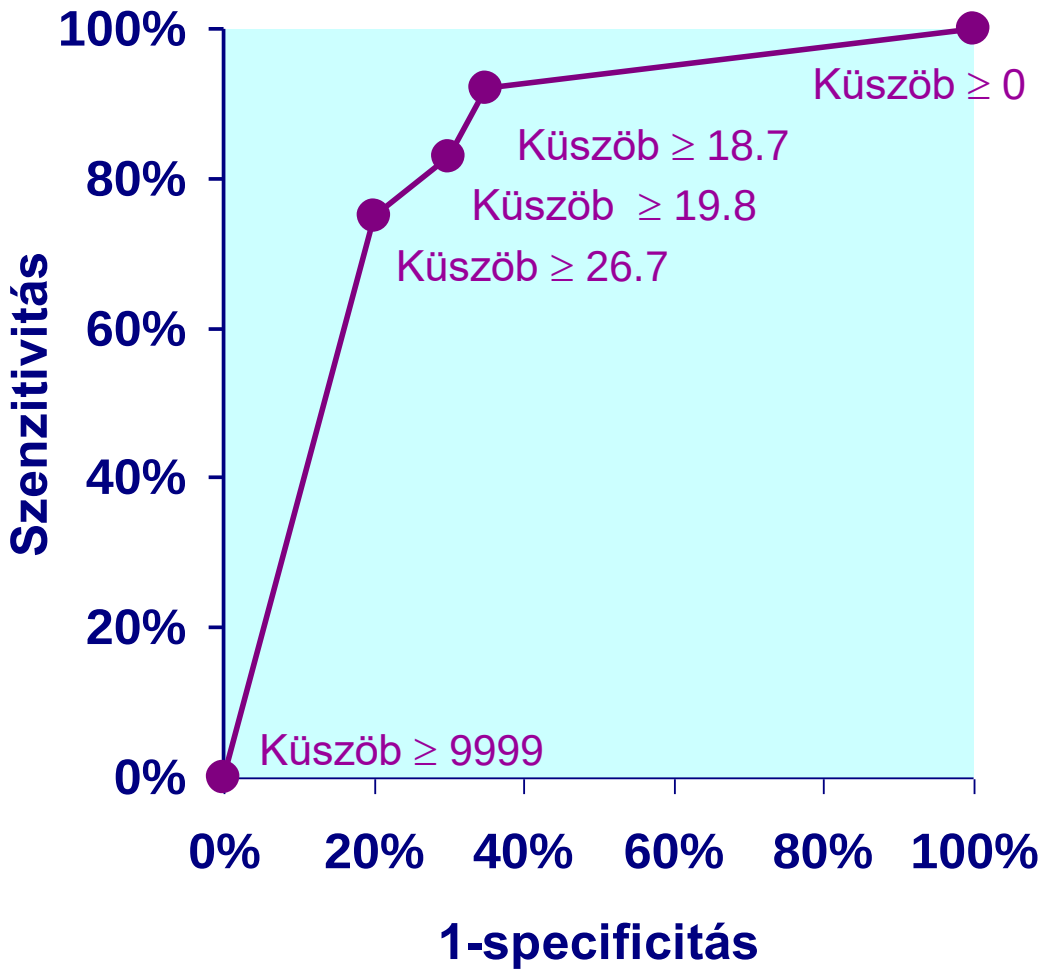
Küszöb	Szenz.	Spec.
≥ 9999	0%	100%
≥ 26.7	75%	80%
≥ 19.8	83%	70%
≥ 18.7	92%	65%
≥ 0	100%	0%



Küszöb	Szenz.	Spec.
≥ 9999	0%	100%
≥ 26.7	75%	80%
≥ 19.8	83%	70%
≥ 18.7	92%	65%
≥ 0	100%	0%

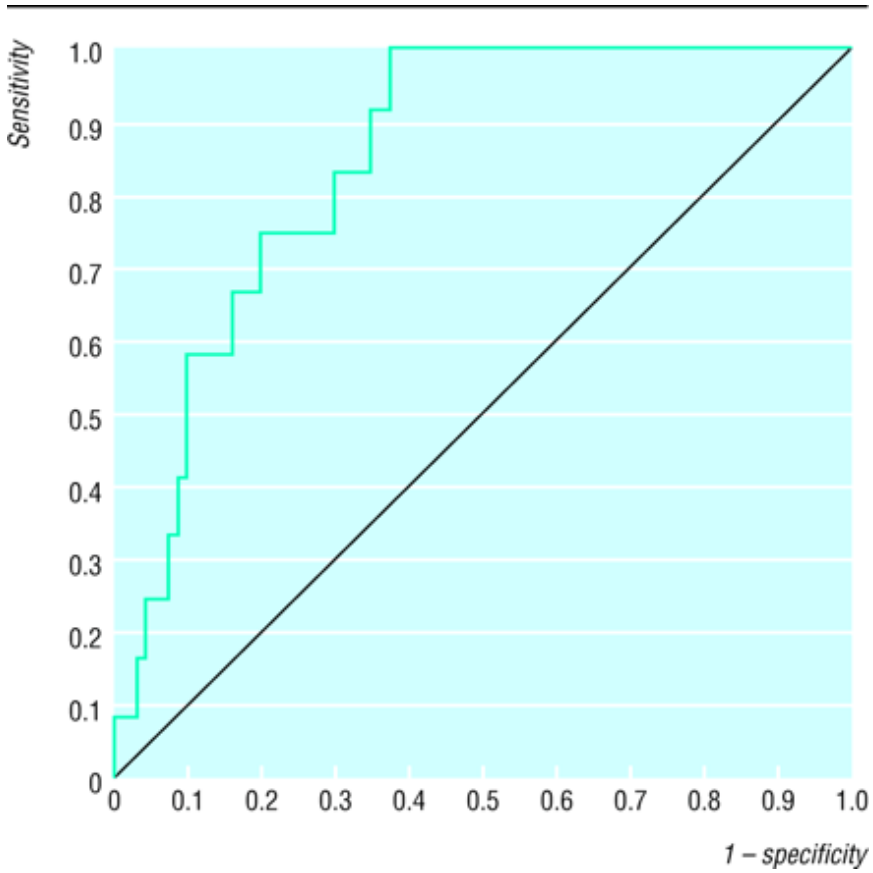


ROC-görbe a BNP-re



Küszöb	Szenz.	Spec.	1-spec.
≥ 9999	0%	100%	0%
≥ 26.7	75%	80%	20%
≥ 19.8	83%	70%	30%
≥ 18.7	92%	65%	35%
≥ 0	100%	0%	100%

ROC görbe több osztópont alapján (a cikk ábrája)



A diagnosztikus küszöb-érték változtatása

Hogyan lehet a legjobb küszöböt megválasztani? Mert ettől függ az összes mérőszám.

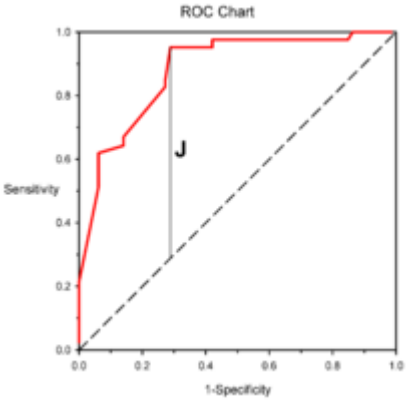
Függ még

- *a betegség prevalenciájától* (amely a PPV-t és NPV-t is befolyásolja)
- az álnegatív és álpozitív félrediagnosztizálás relatív *költségétől*
- az valódi pozitív és a valódi negatív diagnózis relatív *hasznától*

Optimális küszöb-érték megválasztása

- Érdemes pont-ábrát készíteni az elválasztandó pontokról. Orvosi, gazdasági és egyéb szempontokat is figyelembe kell venni.
- Egy leggyakrabban alkalmazott elválasztópont: az a hely, ahol az ún. Youden index a legnagyobb.
  Youden-index=szenzitivitás + specificitás -1

Küszöb	Szenz	Spec	Összeg
100	0%	100%	100%
26,70	75%	80%	155%
19,80	83%	70%	153%
18,70	**92%**	**65%**	**157%**
0,00	100%	0%	100%



https://en.wikipedia.org/wiki/Youden's_J_statistic