# Kvantitativne metode u psihologiji

## KLASTER ANALIZA ILI ANALIZA GRUPISANJA

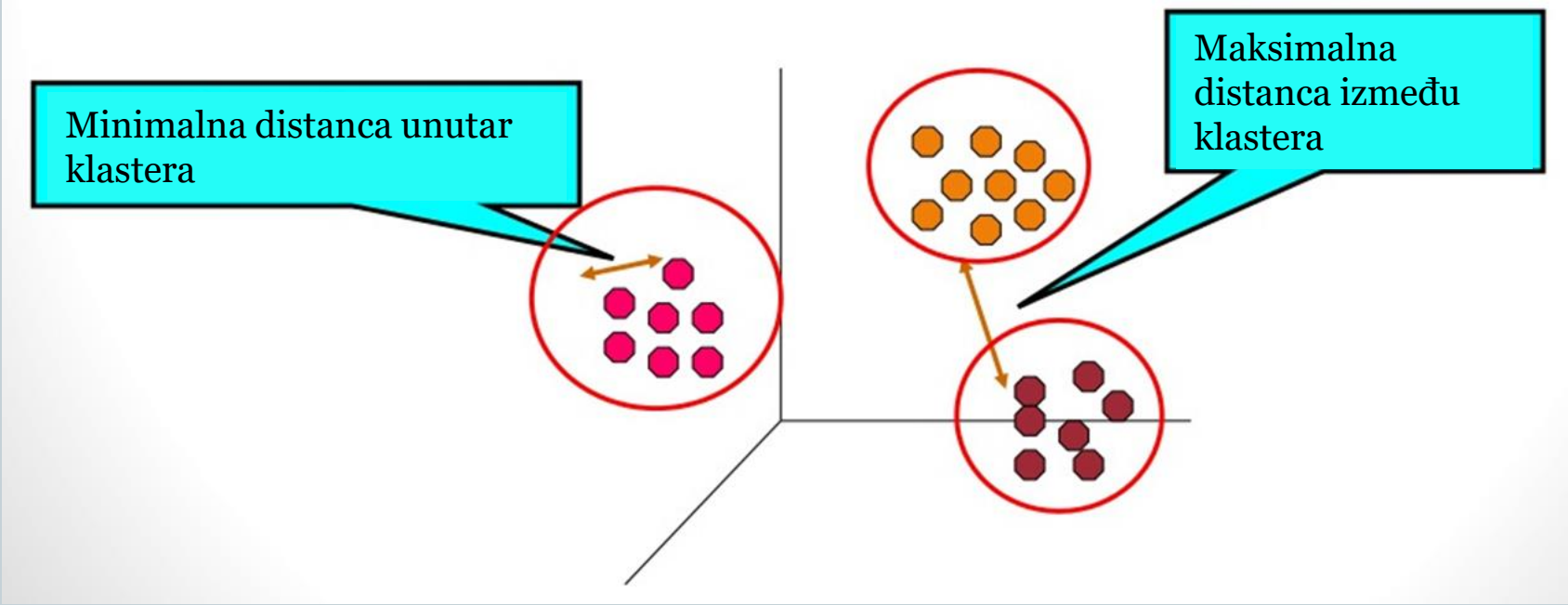
# Šta je klaster analiza?

- Metod multivarijacione analize koji se koristi za grupisanje objekata u grupe, tako da su objekti unutar grupe sličniji međusobom, a između grupa znatno različiti.

- Klaster je grupa sličnih objekata (ispitanika, opservacija, primjera, članova, pacijenata, lokacija itd)

# Ideja klaster analize

- Homogenost unutar grupa
- Heterogenost između grupa



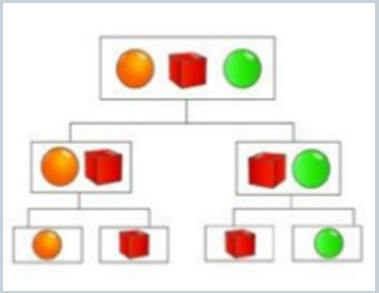
# Neke primjene klaster analize

- Medicina - koji su dijagnostički klasteri? Na bazi ispitanih karakteristika  (anksioznost, depresija itd.) mogu se identifikovati grupe pacijenata koji imaju slične simptome.
- Marketing – na bazi ankete koja pokriva potrebe, stavove, demografiju i ponašanje kupaca, istraživač može koristiti klaster analizu da identifikuje homogene grupa kupaca koji imaju slične potrebe i stavove.
- Obrazovanje - na bazi karakteristika studenata u vezi sa psihološkim i naučnim sposobnostima mogu se identifikovati homogene grupe među studentima (na primjer, visoki uspjesi u svim predmetima ili studenti koji se posebno usavršavaju u pojedinim predmetima, a ne u drugima).
- Ekonomija – grupisanje klijenata u bankama prema kreditnoj istoriji radi ocjene kreditnog rizika.

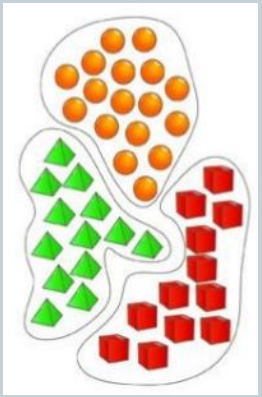
# Tipovi klastera / Metode grupisanja

- Skup pravila pridruživanja objekata u grupe na osnovu mjera bliskosti između objekata

- Hijerarhijske metode – u svakoj iteraciji objekti se pridružuju prethodno formiranim grupama ili sa drugim objektom formiraju novu grupu (hijerarhijsko drvo)



vs



- Nehijerarhijske metode - podjela objekata na podskupove bez preklapanja, tako da se svaki objekt nalazi u tačno jednom klasteru

# Ciljevi klaster analize

- Istraživanje podataka (za otkrivanje nepoznate strukture objekata)
- Redukcija podataka
- Generisanje hipoteza
- Predviđanje

# Klaster analiza vs faktorska i diskriminaciona analiza

- Klaster analiza – grupisanje na bazi distance (bliskosti)
- Faktorska analiza – grupisanje na bazi obrasca varijacije (korelacije)
- Za razliku od faktorske analize, klaster analiza vrši redukciju podataka s obzirom na broj objekata, a ne s obzirom na broj promjenljivih.
- Kod diskriminacione analize grupe su unaprijed poznate, a kod klaster analize ne znamo broj grupa, kao ni šta/ko pripada kojoj grupi.

# Najčešće kritike klaster analize

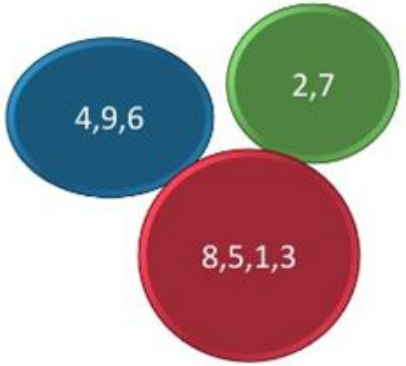
- Klaster analiza je deskriptivna, ateorijska i eksplanatorna.
- Klaster analizom će se uvijek dobiti klasteri bez obzira da li stvarno postoji struktura u podacima.
- Rješenje dobijeno klaster analizom se ne može generalizovati jer je u potpunosti zavisno od korišćenih varijabli na osnovu mjera bliskosti.

# Primjer

| | Maths | Science | Gk | Apt |
|---|---|---|---|---|
| Student-1 | 94 | 82 | 87 | 89 |
| Student-2 | 46 | 67 | 33 | 72 |
| Student-3 | 98 | 97 | 93 | 100 |
| Student-4 | 14 | 5 | 7 | 24 |
| Student-5 | 86 | 97 | 95 | 95 |
| Student-6 | 34 | 32 | 75 | 66 |
| Student-7 | 69 | 44 | 59 | 55 |
| Student-8 | 85 | 90 | 96 | 89 |
| Student-9 | 24 | 26 | 15 | 22 |

| | Maths | Science | Gk | Apt |
|---|---|---|---|---|
| Student-1 | ✓ 94 | ✓ 82 | ✓ 87 | ✓ 89 |
| Student-2 | ! 46 | ! 67 | ✗ 33 | ✓ 72 |
| Student-3 | ✓ 98 | ✓ 97 | ✓ 93 | ✓ 100 |
| Student-4 | ✗ 14 | ✗ 5 | ✗ 7 | ✗ 24 |
| Student-5 | ✓ 86 | ✓ 97 | ✓ 95 | ✓ 95 |
| Student-6 | ✗ 34 | ✗ 32 | ✓ 75 | ! 66 |
| Student-7 | ✓ 69 | ! 44 | ! 59 | ! 55 |
| Student-8 | ✓ 85 | ✓ 90 | ✓ 96 | ✓ 89 |
| Student-9 | ✗ 24 | ✗ 26 | ✗ 15 | ✗ 22 |

| | Maths | Science | Gk | Apt |
|---|---|---|---|---|
| Student-4 | ✗ 14 | ✗ 5 | ✗ 7 | ✗ 24 |
| Student-9 | ✗ 24 | ✗ 26 | ✗ 15 | ✗ 22 |
| Student-6 | ✗ 34 | ✗ 32 | ✓ 75 | ! 66 |
| Student-2 | ! 46 | ! 67 | ✗ 33 | ✓ 72 |
| Student-7 | ✓ 69 | ! 44 | ! 59 | ! 55 |
| Student-8 | ✓ 85 | ✓ 90 | ✓ 96 | ✓ 89 |
| Student-5 | ✓ 86 | ✓ 97 | ✓ 95 | ✓ 95 |
| Student-1 | ✓ 94 | ✓ 82 | ✓ 87 | ✓ 89 |
| Student-3 | ✓ 98 | ✓ 97 | ✓ 93 | ✓ 100 |



# Formiranje klastera

- Odabrati **mjere udaljenosti** (bliskosti)
- Odabrati **metodu klasterizacije**
- Utvrditi **distancu između klastera**
- Odrediti **broj klastera**
- **Validirati** analizu

- Cilj je da podijelimo cijelu populacije u grupe sa sličnim objektima

# Mjerenje sličnosti ili razlike između grupa

- **Menhetn udaljenost –** udaljenost dvije tačke je suma apsolutnih razlika njihovih koordinata na odgovarajućim dimenzijama
- **Euklidska distanca –** ono što obično nazivamo daljinom (u dvodimenzionalnim uslovima to zovemo »vazdušna linija«)
- **Kvadrirana euklidska distanca –**prethodno na kvadrat
- **Mahalonubisova udaljenost –** kvadrirana euklidska distanca korigovana za koreliranost između dimenzija
- **Udaljenost Minkowskog –** bilo koja distanca koja se računa tako što se razlike između koordinata na svakoj od dimenzija podignu na neki stepen, saberu, a potom se iz zbira izvuče koren jednak stepenu na koji su podizane razlike.
- **Chebychev –** distanca dva klastera je razlika između vrijednosti dva klastera na varijabli na kojoj se ta dva klastera najviše razlikuju.

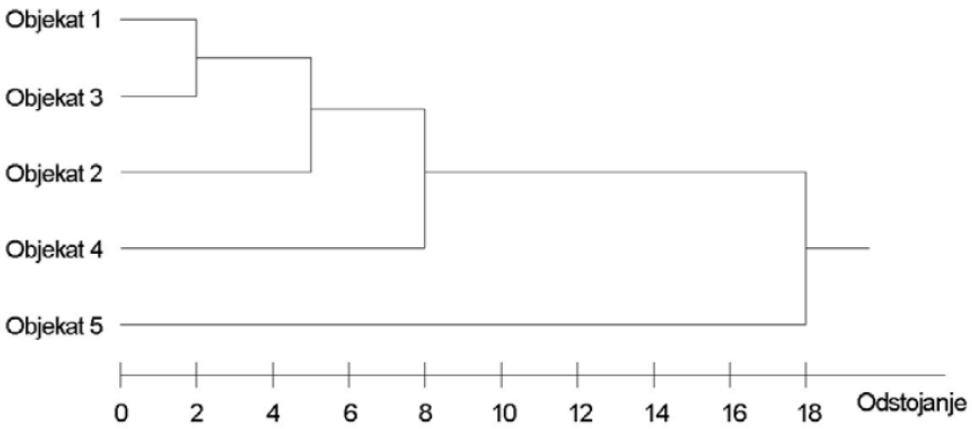
$$D_{ij} = \sum_{k=1}^{n} \left| x_{ki} - x_{kj} \right|$$

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left( x_{ki} - x_{kj} \right)^2}$$

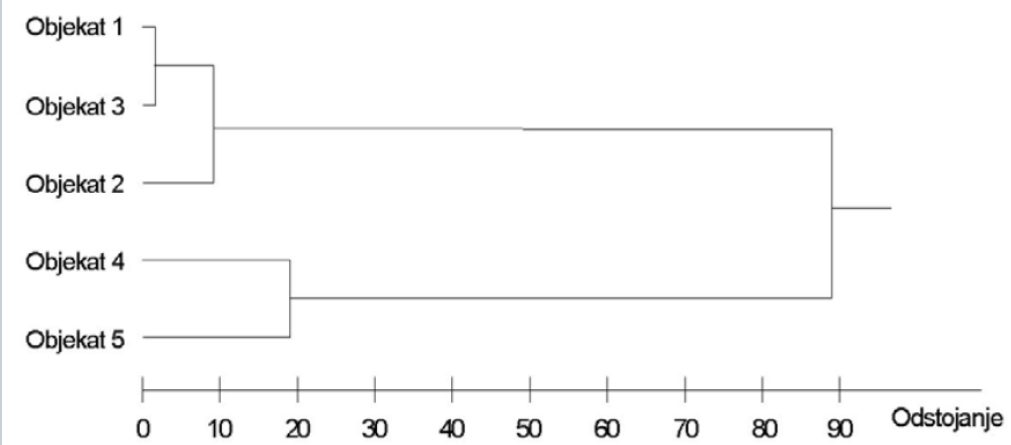
# Metode klasterizacije (hijerarhijske)

- **Metoda najbližih susjeda (Nearest neighbour) –** računa se udaljenost dva najbliža entiteta iz dva klastera
- **Metoda najdaljih susjeda (furthest neighbour) –**uzimaju se dva najdalja entiteta
- **Metoda međugrupog povezivanja (Between-group linkage) –** udaljenost dva klastera je prosjek udaljenosti entiteta iz dva klastera
- **Metoda unutargrupnog povezivanja (Within-group linkage) –** udaljenost dva klastera je prosjek udaljenosti svih parova koji se dobiju kada se ta dva klastera spoje u jedan
- **Centroidna metoda (Centroid clustering) –** udaljenost dva klastera je udaljenost njihovih centroida (centroid je tačka koja predstavlja atitmetičku srednu (prosjek) položaja svih tačaka unutar određenog područja)
- **Metoda medijane (Median Clustering) –** isto kao prethodna samo što nema pondera, oba klastera podjednako doprinose vrijednosti centroida
- **Vordova metoda (Ward's Method) –** računaju se sume udaljenosti entiteta od centra klastera, a klasterizacija se radi tako da se ove sume u klasterima minimizuju

# Dendogram – grafički prikaz hijerarhijske strukture




**Slika 9.5** Dendrogram jednostrukog povezivanja




**Slika 9.6** Dendrogram potpunog povezivanja

# Izbor broja grupa

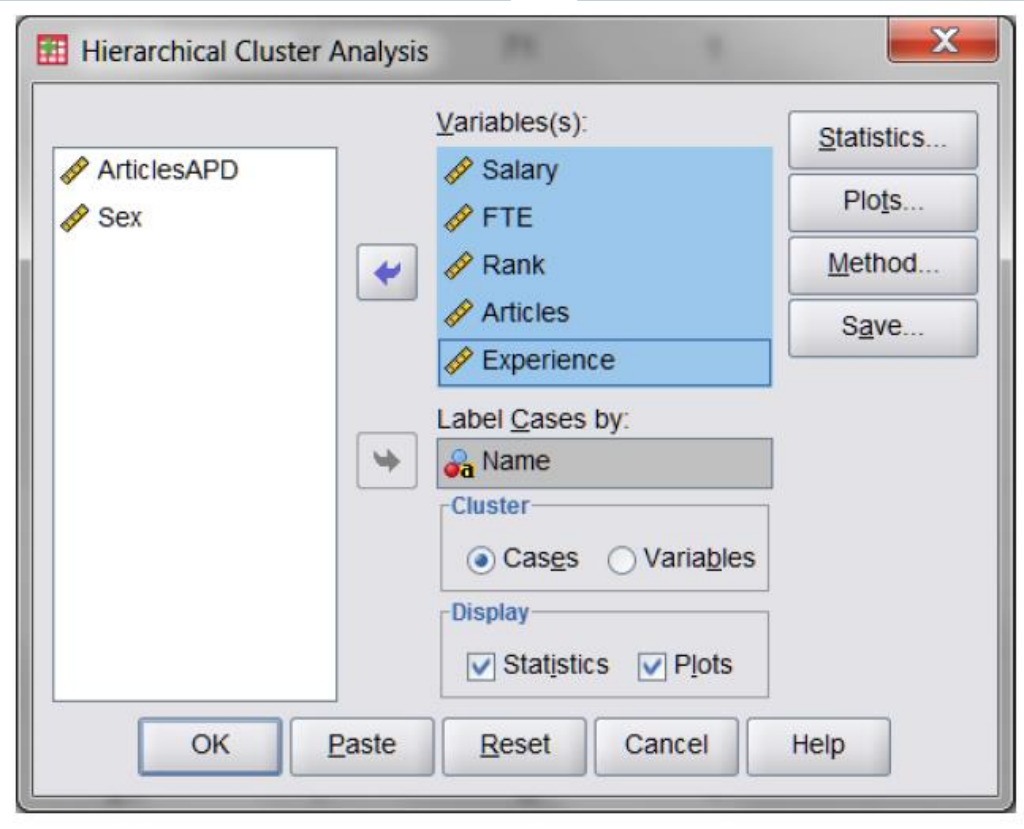
- Najjednostavniji pristup – izbor broja grupa zasnovan na praćenju vrijednosti mjera odstojanja.
- Krećući se od prvog koraka, vrijednost mjere raste. Ako se u određenom koraku zabilježi velika promjena u vrijednosti te mjere odstojanja između grupa, tada broj grupa koji prethodi tom koraku proglašavamo optimalnim.

| Korak | Broj grupa | Grupe | Odstojanje | Priraštaj odstojanja |
|---|---|---|---|---|
| 1 | 9 | (Marko) (Miodrag) | 0.010 | - |
| 2 | 8 | (Stojan) (Ljubomir) | 0.040 | 0.030 |
| 3 | 7 | (Vladan) (Bojan) | 0.040 | 0 |
| 4 | 6 | (Stojan, Ljubomir) (Milena) | 0.060 | 0.020 |
| 5 | 5 | (Marko, Miodrag) (Danica) | 0.065 | 0.005 |
| 6 | 4 | (Vladan, Bojan) (Zorica) | 0.110 | 0.045 |
| 7 | 3 | (Marko, Miodrag, Danica) (Dubravka) | 0.160 | 0.050 |
| 8 | 2 | (Stojan, Ljubomir, Milena) (Marko, Miodrag, Danica, Dubravka) | 0.832 | 0.678 |
| 9 | 1 | (Stojan, Ljubomir, Milena, Marko, Miodrag, Danica, Dubravka) (Vladan, Bojan, Zorica) | 1.878 | 1.040 |

# Nehijerarhijske metode klasterizacije

- Pretpostavlja da se da je broj grupa unaprijed poznat ili ga variramo tokom postupka grupisanja
- Postupak počinje inicijalnom podjelom skupa podataka u izabran broj grupa
- Zatim se određuje odstojanje između svakog objekta i svake grupe
- Nakon pridruživanja objekata izračunava se centroid grupe iz koje je objekat otišao i grupe kojoj se pridružio
- Zatim se za svaki objekat ponovo računaju odstojanja od centroida grupa
- Najpopularniji nehijerarhijski metod je metod k-sredina (K-means method)
- Metodi grupisanja osjetljivi na prisustvo nestandarnih opservacija

# Primjeri u SPSS-u




SPSS: Analyze/Classify/Hierarchical Cluster...