Uvod u multivarijacionu analizu

Koncept multivarijacione analize

Multivarijaciona analiza (MA) se pojavila u nauci pre približno jednog veka. Ozbiljnija primena u ekonomiji je bila u oblasti marketinga u ranim pedesetim godinama XX veka i od tada njena upotreba se širi i na ostale grane ekonomije, a takođe i na druge naučne oblasti. Na početku su sva izračunavanja, koja su bila veoma složena, izvođena ručno, a pojavom modernih računara i softvera omogućeno je da se sa par „klikova" izvede analiza za koju je nekad trebalo po nekoliko sati. Ovaj tehnološki napredak je doprineo još većoj primeni i popularnosti MA.

Pre pojave MA velika većina analize podataka je podrazumevala analizu jedne ili najviše dve varijable istovremeno. Varijable su bili prikazane u prostoj ili složenoj (ukrštenoj) tabeli u slučaju dve varijable (cross-tabulated). Analiza je podrazumevala izračunavanje sledećih statistika iz uzorka:

- mere centralne tendencije
- mere varijacije
- intervale poverenja
- testiranje hipoteza o parametrima osnovnog skupa
- koefijent korelacije.

MA pruža mnogo bolje alate koji omogućavaju otkrivanje zakonitosti u odnosima varijabli koje su skrivene ili jedva primetne. Osim toga, većina tehnika je dovoljno precizna da se preko testiranja statističke značajnosti utvrdi da li su otkrivene zakonitosti značajne ili slučajne, odnosno plod slučajnih varijacija podataka u uzorku. Primenom MA povećava se obim relevantnih informacija koje se mogu „izvući" iz nekog skupa podataka. „Dobra analogija je kontrast između slike u dve boje i iste te slike u punom koloru." (Myers & Mullet, 2003, str. 4).

Predmet MA su varijable (promenljive). Varijabla je zapravo skup varijacija nekog obeležja u uzorku ili osnovnom skupu. Na primer, visina ličnog dohotka radnika, broj prodatih proizvoda po prodavnicama, stopa nezaposlenosti po regionima, BDP po zemljama itd.

Pojam „multivarijacioni" jednostavno podrazumeva da je u pitanju veliki broj varijabli. Postoji neslaganje u domaćoj literaturi koji pojam najbolje odgovara originalnom nazivu na engleskom koji glasi „multivariate analysis". Koriste se sledeći izrazi: „multivarijaciona", „multivarijantna", pa čak i „multivarijatna" analiza. Možda bi zapravo pravi termin bio „multivarijabilna analiza" ako je već reč o velikom broju varijabli.

Multivarijacione tehnike se koriste za simultanu analizu međusobnog odnosa između velikog broja varijabli (više od dve) na bazi određenog modela na kojem se zasniva data tehnika. Većina tehnika identifikuje šablone ili sličnosti u odnosima između varijabli i na osnovu toga vrši objašnjavanje tog odnosa ili čak njegovo predviđanje. Inače, MA ima interdisciplinarni karakter jer je ona nastala na osnovu rada velikog broja stručnjaka iz različitih naučnih oblasti a i njena primena (i mogućnost primene) je toliko široka da je lakše nabrojati one oblasti u kojima ona nema svoju primenu.

Tehnike MA mogu da se podele u dve grupe: tehnike zavisnosti i tehnike međuzavisnosti.

Kod tehnika zavisnosti postoji specifičan odnos između varijabli gde na jednoj strani imamo zavisne (rezultujuće) varijable na koje uticaj vrše nezavisne varijable. U ovu grupu spadaju sledeće tehnike:

- Višestruka regresija i korelacija
- Logit analiza
- Diskriminaciona analiza
- Kanonička analiza
- Conjoint analiza
- AID, CHAID i CART analiza

Sa druge strane, tehnike međuzavisnosti polaze od podjednake važnosti i uloge svih varijabli i sve se posmatraju ka nezavisne. Kod ovih tehnika primarni cilj je utvrditi odnos između njih, najčešće sa željom da se pronađu grupe varijabli koje su slične u svojim varijacijama i da pri tome njihovo grupisanje ima određeni smisao. Na primer, kada pri segmentaciji nekog tržišta identifikujemo skupove potrošača koji imaju niz zajedničkih karakteristika (varijabli) odnosno ponašaju se na isti način na tržištu. U ovu grupu spadaju sledeće tehnike:

- Analiza glavnih komponenti
- Faktorska analiza
- Klaster analiza
- Korespondentna analiza
- Strukturalne jednačine
- Multidimenzionalno skaliranje

U praksi često dolazi do kombinovanja više tehnika. Na primer, na osnovu klaster analize izvrši se segmentacija tržišta a zatim se na osnovu regresione i korelacione analize traže varijable koje imaju najvećeg uticaja na svakom od segmenata.

Već je pomenuto da se MA danas radi uz pomoć informacionih tehnologija. Svi statistički malo kvalitetniji statistički softveri imaju opcije koje omogućavaju da se izvrše ako ne baš sve, a ono barem najčešće korišćene multivarijacione tehnike. Kada se jednom formira baza podataka, sa izborom svega nekoliko opcija postaje moguće izvesti i najsloženije tehnike za par sekundi.

Još jedna od važnih činjenica jeste da su računari omogućili da se veliki broj grafičkih tehnika može primeniti u multivarijacionoj analizi. U pitanju su grafički prikazi kontura, trodimenzionalni dijagrami i čitav spektar specijalnih grafičkih prikaza od kojih će neki biti prikazani i ovde.

„Dobra vest je da čak i neistrenirani i neiskusni ljudi mogu da izvedu komplikovane multivarijacine analize uz pomoć računara. Loša vest je da čak i neistranirani i neiskusni ljudi mogu da izvedu komplikovane multivarijacione analize uz pomoć računara!“ (Myers & Mullet, 2003, str. 11).

Multivarijacioni podaci

Na osnovnim kursevima iz statistike akcenat je na univarijacionim metodima koji se bave analizom varijacija jedne slučajne promenljive. Kada je u pitanju multivarijaciona analiza, posmatra se više povezanih varijabli istovremeno, gde na početku analize imaju isti značaj. U nastavku sledi prezentacija jednog multivarijacionog sistema podataka.

Jedinice posmatranja	Varijable			
	X_1	X_2	...	X_m
1.	a_{11}	a_{12}	...	a_{1m}
2.	a_{21}	a_{22}	...	a_{2m}
...	..	...	...	...
n	a_{n1}	a_{n2}	...	a_{nm}

Primer: Stope aktivnosti u zemljama jugoistočne Evrope, prema polu, 2002. godine.

Posmatrajući dati set podataka, javljaju se sledeći problemi koji se mogu rešavati primenom multivarijacione analize:

- Kako su postojeće varijable povezane? Koliko jako su povezane?
- Da li je rast neke varijable povezan sa opadanjem druge?
- Da li se jedinice posmatranja statističkih značajno razlikuju u pogledu vrednosti varijabli?

- Da li jedinice posmatranja na isti način variraju od varijable do varijable?
- Da li je moguće iz velikog broja varijabli izdvojiti one koje su najznačajnije i koje bi reprezentovale ceo skup varijabli, odnosno da li se mogu otkriti dimenizije koje objedinjavaju varijacije više varijabli zajedno?
- Koje jedinice posmatranja su međusobno najsličnije a koje se najviše razlikuju?
- Da li možemo da skup jedinica posmatranja podelimo u podskupove na osnovu postojeći varijacija?
- Da li je moguće za novu jedinicu posmatranja na osnovu samo dve-tri vrednosti varijabli odrediti u koju grupu spada?

U zavisnosti od konkretnog slučaja, pojaviće se samo neka od navedenih pitanja (ili možda neka druga), i odabraće se ona multivarijaciona tehnika koja može da reši odgovarajući problem.

Osnovni pojmovi

Modeli i heuristike. Multivarijacione tehnike sa aspekta preciznosti delimo na modele i heuristike. Modeli imaju veću preciznost, mogu se u potpunosti definisati preko simbola i obično imaju formalnu matematičku derivaciju. Ukoliko su sve teorijske pretpostavke zadovoljene, modeli daju tačno i optimalno rešenje. U MA modeli se više koriste kod tehnika zavisnosti.

Heuristike se zasnivaju na procedurama traženja i pružaju rešenja kroz iteracije odnosno ponavljanja procedure, gde se sa svakom iteracijom približavamo optimalnom rešenju. Ipak, kod ovih tehnika ne postoji matematički optimizacioni model tako da analitičar treba da donese subjektivnu ocenu o tome koje rešenje je za njega optimalno, odnosno kada treba prestati sa iteracijama.

Modeli. Modeli su formalne prezentacije realnog sveta. Matematički modeli obično imaju tri komponente: varijable, konstante i relacije. Varijable su obeležja koja mogu da variraju na način koji može da se identifikuje ili izmeri. Konstante su fiksne vrednosti koje po definiciji ne variraju i delimo ih na one koje stoje same i na one koje stoje uz varijable i zovu se koeficijenti. Koeficijenti pokazuju koliko jako varijabla utiče na ishod modela. Relacije pokazuju međusobni odnos varijabli i konstanti i predstavljaju se simbolima za matematičke operacije.

Na primer, sledeća jednačina predstavlja jedan model:

Matematički modeli ne moraju da budu samo jednačine, ali ovde će se samo one koristiti za predstavljanje modela. Navedeni model se naziva „bivarijantni“ jer se pojavljuju dve varijable (X i Y), za razliku od modela koji imaju više varijabli pa se nazivaju „multivarijacioni“. Na primer:

Multivarijacioni modeli imaju sledeće osobine:

- Oni su pre jednačine nego nejednačine.
- Oni su pre linearni nego nelinearni.
- Oni su aditivni jer se nezavisne varijable pre dodaju ili oduzimaju jedna od druge nego što se množe ili dele.
- Oni su kompenzatorni jer niska vrednost jedne nezavisne varijable može da bude kompenzovana visokim vrednostima drugih nezavisnih varijabli.

Sa aspekta preciznosti razlikuju se deterministički i probabilistički (stohastički) modeli. Deterministički modeli pokazuju precizan odnos između svih varijabli u modelu i oni se definišu pre svega u prirodnim naukama. Probabilistički modeli pokazuju odnos između varijabli na osnovu njihovih najverovatnijih vrednosti. Oni su tipični za pojave iz oblasti društvenih nauka gde je prisutna varijabilnost i nepredvidljivost varijabli. Deterministički model može da ima sledeći oblik:

Kod probabilističkog modela potrebno je uvrstiti i grešku *e* koja govori o „nesavršenosti“ datog modela:

Merne skale. Da bi se neka multivarijaciona tehnika mogla primeniti, potrebno je znati koja merna skala je korišćena za svaku varijablu. Brojčane vrednosti se ne mogu razumeti i interpretirati na pravi način ako se ne zna na osnovu koje merne skale su izražene. Postoje četiri osnovna tipa mernih skala: nominalna, ordinalna, intervalna i racio.

Nominalna skala je ona kod koje ne postoji redosled u vrednostima s obzirom na njihovu veličinu. To su prosto imena odnosno nazivi za određene kategorije koje ne stoje ni na kakvoj vrednosnoj skali i služe samo za identifikaciju. Obično se za ove podatke kaže da su kategorički. Na primer, nazivi gradova, reka ili imena ljudi predstavljaju vrednosti na nominalnoj skali. Takođe, kada na dresovima sportista stoji broj i to se smatra nominalnom vrednošću jer ako neki igrač nosi broj „8“ na dresu a drugi igrač broj „4“ to ne znači da je prvi igrač dva puta bolji ili vredniji od drugog.

Ordinalna skala se koristi onda kada brojevi stoje u sekvenci koja ukazuje na redosled ili veličinu. Često se koristi i izraz „rang“. Na primer, rang 1 najčešće predstavlja najveću vrednost ili poziciju, rang 2 nešto manju, rang 3 još manju itd. Takođe, može da bude i obrnuto. Međutim, razlika u veličini između susednih vrednosti može da bude različita jer rangovi ne daju informaciju o tim razlikama. Na primer, u nekom sportu, na tabeli najbolji tim dobija rang 1, sledeći rang 2 itd. ali to ne znači da je ista razlika između prvog i drugog tima kao što je između drugog i trećeg, trećeg i četvrtog itd.

Zbog toga što ne daju mernu vrednost, za nominalne i ordinalne skale kažemo da su „nemetričke“ i za njihovu analizu se uglavnom koriste neparametarski testovi. Ipak, u društvenim naukama često imamo posla baš sa ovakvim vrednostima.

Intervalne skale pokazuju da između svih susednih vrednosti na njima postoji jednak razmak. Često se koristi termin „kadinalna vrednost“ da bi se ukazalo da su u pitanju intervalne vrednosti koje su jednake međusobno. To zapravo znači da je razlika između, recimo, 3 i 4 ista kao i razlika između 8 i 9 ili 16 i 17. Na ovakvim skalama ne postoji apsolutna nula koja bi ukazivala na nepostojanje vrednosti koja se meri. Iz ovakvih vrednosti ne treba računati procente ili razmere. Na primer, intervalne skale postoje kod merenja inteligencije, znanja, interesovanja, ukusa itd. Ovakve skale se često koriste u marketingu da bi se izmerili stavovi potrošača, percepcije, interesi itd. Kao primer može da posluži temperatura: Ako je u jednom danu izmereno 20 celzijusa a u drugom 40, to ne znači da je drugog dana bilo duplo toplije. Ova skala ima nulu ali ona je dodeljena arbitrarno i tu nula ne odslikava odsustvo temperature.

Racio skale. I kod ovih skala postoje jednaki intervali između susednih vrednosti, sa tom razlikom što postoji apsolutna nula. To nam omogućava da izračunamo relativne promene ili razlike (procente). Najveći broj numeričkih obeležja kod ekonomskih pojava, i prekidnih i neprekidnih, spada u ovu grupu.

Intervalne i racio skale se nazivaju još i metričkim skalama i za njihovu analizu se najčešće koriste parametarski testovi.

Testiranje validnosti i pouzdanosti merenja

Multivarijaciona normalna distribucija

Za razumevanje materije o multivarijacionoj analizi neophodno je poznavanje normalnog rasporeda (distribucije) za jednu varijablu. Poznato je da ta distribucija frekvencija ima oblik zvona i da su mnogi univarijacioni statistički metodi bazirani na pretpostavci da originalni podaci imaju normalni raspored.

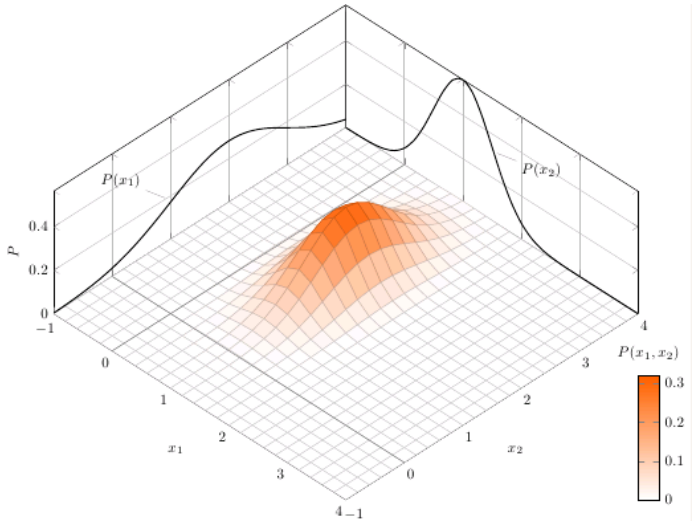
Za očekivati je da kod multivarijacionih statističkih tehnika multivarijaciona normalna distribucija ima centralnu ulogu. Mnoge tehnike se takođe zasnivaju na pretpostavci da originalni podaci imaju multivarijacionu normalnu distribuciju.

Tačna definicija ove distribucije i njen matematički izraz nije neophodan za primenu multivarijacionih tehnika. Najčešći pristup jeste da se originalni podaci tretiraju kao normalno distribuirani sve dok ne postoji neki jak razlog da to možda

nije slučaj. Takođe, ako pojedinačne varijable imaju normalnu distribuciju, onda se pretpostavlja da i objedinjena distribucija svih varijabli ima normalan raspored.

Može da se desi da bude očigledno da multivarijacioni podaci nemaju normalan raspored jer je prisutna jaka asimetrija u jednu stranu sa nekoliko veoma velikih ili malih vrednosti. Ovaj problem može da se reši odgovarajućom transformacijom podataka ili primenom specijalnih tehnika za analizu.

Veoma važna karakteristika multivarijacione distribucije jeste da je ona u potpunosti definicana vektorom srednjih vrednosti i matricom kovarijansi kao što je i univarijaciona normalna distribucija određena aritmetičkom sredinom i varijansom. Vektor srednjih vrednosti sadrži aritmetičke sredine svih varijabli dok matrica kovarijansi sadrži varijanse svih varijabli plus kovarijanse, koje pokazuju koliko su parovi varijabli povezani.



Izvor: http://tex.stackexchange.com/questions/31708/draw-a-bivariate-normal-distribution-in-tikz. Datum preuzimanja: 27.11.2012.

Matrična algebra

U osnovi računskih operacija vezanih za multivarijacionu analizu leži upotreba matrične algebre. Veći deo matrične algebre se izučava na osnovnom nivou iz matematike i nekih drugih kvantitativnih disciplina pa će ovde biti nabrojani oni osnovni pojmovi koji će se koristiti u daljem tekstu:

- matrice
- vektori
- kvadratna matrica
- transponovanje matrice
- nula matrica
- dijagonalna matrica
- simetrična matrica
- matrica identiteta (jedinična matrica)
- skalari i množenje skalarom
- osnovne četiri računske operacije sa matricama
- inverzne matrice
- determinante
- ortogonalne matrice

Kvadratna forma. Neka je **A** matrica *n* x *n* i **x** je vektor kolona dužine *n*. Onda je kvantitet

skalar koji se zove kvadratna forma. Ovaj skalar se može predstaviti i na sledeći način:

gde je x_i elemenat *i*-tog reda vektor kolone **x** a a_{ij} je elemenat *i*-tog reda i *j*-te kolone matrice *A*.

Ajgenvrednosti i ajgenvektori. Dat je sledeći set linearnih jednačina:

...

gde je λ skalar. Prethodni sistem jednačina može da se napiše i u matričnoj formi:

ili

gde je **I** jedinična matrica *n* x *n*, a **0** je nula vektor *n* x 1. Može se dokazati da date jednačine važe samo za određene vrednosti λ koje se zovu latentni koreni ili ajgenvrednosti matrice **A**. Može da bude do *n* takvih ajgenvrednosti. Za datu *i*-tu vrednost λ_i, jednačine se mogu rešiti kada se odredi da je x_1=1, i rezultujući vektor **x** vrednosti se transponuje u **x'**= (1, x_2, x_3,...,x_n) ili neki proizvod ovog vektora, zove se *i*-ti latnentni koren ili *i*-ti ajgenvektor matrice **A**. Takođe, suma ajgenvrednosti je jednaka sa zbirom elemenata na glavnoj dijagonali kvadratne matrice (trace).

Zbir elemenata na glavnoj dijagonali (**A**) = $\lambda_1+\lambda_2+...+\lambda_n$.

Vektori srednjih vrednosti i matrice kovarijansi.

Vrednosti u osnovnom skupu ili uzorku jedne slučajne varijable su često predstavljeni preko aritmetičke sredine i varijanse. Aritmetička sredina i varijansa uzorka se izračunavaju na sledeći način:

Ovo su ocene koje odgovaraju parametrima u osnovnom skupu, odnosno aritmetičkoj sredini osnovnog skupa μ i varijansi populacije σ^2.

Na sličan način, multivarijaciona populacija i uzorci mogu da budu predstavljeni vektorima srednjih vrednosti i matricama kovarijansi. Neka ima ukupno *p* varijabli, sa uzorkom od n vrednosti za svaku varijablu. Neka su aritmetička sredina uzorka i varijansa uzorka za *i*-tu varijablu i i da su izračunate na prethodno prikazani način. Pored toga kovarijansa uzorka između varijabli X_j i X_k je:

gde su x_{ij} vrednosti varijable X_j za *i*-tu multivarijacionu opservaciju. Kovarijansa je mera koja pokazuje kolika je linearna povezanost između dve varijable gde pozitivna vrednost pokazuje da se velike vrednosti jedne varijable poklapaju sa velikim vrednostima druge, a negativna vrednost pokazuje da se velike vrednosti jedne varijable poklapaju sa malim vrednostima druge.

Kovarijansa je povezana sa koeficijentom korelacije između dve varijable:

Takođe, važe sledeće jednakosti:

Matrica varijansi i kovarijansi uzorka ili skraćeno matrica kovarijansi ima sledeći oblik:

Ova matrica se ponekad zove i matrica disperzije uzorka i meri iznos varijacija uzorka kao i stepen korelacije p varijabli. Ona predstavlja ocenu matrice kovarijansi populacije:

Na kraju, matrica korelacije u uzorku je sledeća:

Naravno, ova matrica je ocena odgovarajuće matrice korelacije osnovnog skupa.

Veoma važan rezultat za veliki broj multivarijacionih analiza jesu standardizovane vrednosti varijabli koje se dobijaju tako što se aritmetička sredina uzorka oduzme od originalne vrednosti i podeli sa standardnom devijacijom. Na taj način se dobijaju vrednosti varijable koje imaju aritmetičku sredinu jednaku nuli i standardnu devijaciju jednaku jedinici. U tom slučaju, matrica kovarijansi je jednaka sa matricom korelacije uzorka, $\boldsymbol{C} = \boldsymbol{R}$.

Multivarijacione distance

Veliki broj problema sa multivarijacionim podacima može da se posmatra preko distanci između jedinica posmatranja, između dva ili više uzoraka, ili između dva ili više osnovnih skupova.

Distance između jedinica posmatranja. Neka postoji *n* jedinica posmatranja gde svaka ima vrednosti za *p* broj varijabli (X_1, X_2, ..., X_p). Vrednosti za jedinicu *i* su označene sa x_{i1}, x_{i2}, ..., x_{ip}, a za jedinicu *j* sa x_{j1}, x_{j2}, ..., x_{jp}. Ako su u pitanju samo dve varijable, onda vrednosti mogu da se predstave u dvodimenzionalnom dijagramu i može da se izračuna Euklidova distanca:

Opšti obrazac za izračunavanje Euklidovih distanci kada je u pitanju *p* broj varijabli je sledeći:

Iz ove formule jasno proizilazi da ako jedna od varijabli više varira, da će ona dominirati u izračunavanju distanci, dok će uticaj ostalih varijabli na udaljenost biti umanjen. Zbog toga je poželjno da sve varijable imaju isti uticaj na izračunavanje distanci. Ovo se može postići preko deljenja varijabli sa svojom standardnom devijacijom za *n* jedinica posmatranja koje se upoređuju.

Distance između osnovnih skupova i uzoraka. Postoji veliki broj načina da se izračunaju distance između multivarijacionih populacija kada su poznate aritmetičke sredine, varijanse i kovarijanse osnovnih skupova. Ako pretpostavimo da postoje dva ili više osnovnih skupova za koje je poznata distribucija po p broju varijabli (X_1, X_2, ..., X_p), onda se može izračunati distanca između populacija i i j po formuli Penrose-a (Manly, 2005):

gde je aritmetička sredina osnovnog skupa i za varijablu X_k označena sa μ_{ki}, a pretpostavljamo da je varijansa za X_k u svim osnovnim skupovima V_k. Pomenuti obrazac ima nedostatak jer ne uzima u obzir korelaciju između varijabli. To konkretno znači da ako su dve varijable u visokoj korelaciji, one pojedinačno podjednako doprinose veličini distance između dva osnovna skupa, i to približno podjednako kao i treća varijabla koja nije u korelaciji ni sa jednom drugom varijablom. Trebalo bi, zapravo, da zbog korelacije i obuhvaćenog uticaja na distancu jedne varijable koja je korelisana, uticaj druge korelisane varijable bude umanjem. Zbog toga se često koristi Mahalanobi-jeva distanca koja uzima u obzir korelaciju:

gde je v_{rs} element u r-tom redu i s-toj koloni inverzne matrice kovarijansi osnovnog skupa za p varijabli.

Mahalanobi-jeva distanca se često koristi da bi se izračunala udaljenost jedne jedinice posmatranja od centra osnovnog skupa:

gde su x_1, x_2, ..., x_p vrednosti varijabli X_1, X_2, ..., X_p za datu jedinicu posmatranja, sa odgovarajućim aritmetičkim sredinama μ_1, μ_2, ..., μ_p, $\boldsymbol{x}$ je vektor vrednosti, $\boldsymbol{\mu}$ je vektor aritmetičkih sredina, $\boldsymbol{V}$ je matrica kovarijansi osnovnog skupa, a v_{rs} je element u r-tom redu i s-toj koloni inverzne matrice od $\boldsymbol{V}$.

Ukoliko podaci iz osnovnog skupa imaju multivarijacionu normalnu distribuciju, onda distanca D^2 ima distribuciju hi-kvadrat rasporeda sa p stepeni slobode. Značajno velika vrednost D^2 znači da je jedinica koja se posmatra ili ekstremna vrednost, ili dolazi iz druge populacije, ili je u pitanju neka greška. Drugim rečima, takva jedinica posmatranja mora da se proveri.

Manly (2005) ističe da Mahalanobi distanca ima prednost u odnosu na Penrose-ovu distancu jer koristi informacije o kovarijansama. Ipak, ova prednost je moguća kada su kovarijanse poznate. U slučajevima kada se kovarijanse mogu samo oceniti na osnovu uzoraka, bolje je koristiti Penrose-ovu distancu. Teško je reći tačno šta je u ovom kontekstu mali uzorak, ali se generalno može reći da ako se matrica kovarijansi osnovnog skupa ocenjuje na osnovu uzorka koji ima 100 ili više elemenata, onda se može upotrebiti Mahalanobi distanca.

Distance zasnovane na proporcijama. Specifična situacija koja se javlja kod nekih varijabli jeste da kao mera distanci između populacija ili uzoraka budu upotrebljene proporcije. Na primer, kao jedinice posmatranja uzimaju se opštine, a kao varijable proporcija odnosno procenat radnika zaposlenih u različitim delatnostima. Pitanje koje može da se postavi jeste koliko su međusobno udaljene dve opštine.

Različite distance se koriste u ovakvim slučajevima. Na primer, prva distanca bi bila:

što je polovina sume apsolutnih razlika proporcija. Ovaj indeks ima vrednost jedan ako nema preklapanja između jedinica i nula ako je $p_i=q_i$ za sve vrednosti i. Druga distanca koja se koristi je sledeća:

I u ovom slučaju ako je d_2 bliža jedinici postoji manje preklapanje a ako je bliža nuli proporcije su jednakije.

Poto d_1 i d_2 variraju od nule do jedan, sledi da 1-d_1 i 1-d_2 predstavljaju mere sličnosti između jedinica koje se posmatraju. Koristi se često sledeći izraz:

Ako je s_1=0 onda su dve jedinice (u našem primeru opštine) potpuno različite, a ako je s_1=1 onda su potpuno iste.

Varijable koje pokazuju prisustvo i odsustvo obeležja. Sledeća vrlo česta situacija jeste pojavljivanje varijabli koje imaju samo dve vrednosti (dihotomne varijable): prisustvo (1) i odsustvo (0) datog obeležja. Na primer, postoji interes da se izmeri koliko su slična dva programa zapošljavanja koja su implementirana u 8 opština, odnosno posmatra se da li je dati program implementiran ili nije. Podaci mogu da budu prikazani na sledeći način:

Tablela: Pregled opština prema tome da li su programi zapošljavanja primenjeni ili nisu

Opštine:	1	2	3	4	5	6	7	8
Program 1	1	1	1	0	0	1	0	0
Program 2	1	1	0	0	0	0	1	1

Podaci iz prethodne tabele se mogu sumirati na sledeći način:

Tabela: Prisustvo i odsustvo programa zapošljavanja po opštinama

Program 1	Program 2		
	Prisustvo	Odsustvo	Ukupno
Prisustvo	a	b	$a+b$
Odsustvo	c	d	$c+d$
Ukupno	$a+c$	$b+d$	n

Neke od mera koje se uobičajeno koriste za izračunavanje udaljenosti su sledeće:

Prosti indeks slaganja

Ochiai indeks

Dice-Sorensen indeks

Jaccard indeks

Ovi indeksi variraju od nule (nema sličnosti) do jedinice (potpuna sličnost), tako da se oduzimanjem od jedinice opet mogu dobiti komplementarne mere udaljenosti.

Mantel-ov test randomizacije. Ovo je test koji se koristi za upoređivanje dve matrice udaljenosti i koristi se za rešavanje problema otkrivanja klasterizacije u prostoru i vremenu. Na primer, ako želimo da utvrdimo da li se štrajkovi koji se dešavaju blizu jedan drugom poklapaju i s obzirom na vreme odvijanja.

Pretpostavimo da se tri jedinice posmatranja prate u pogledu dve grupe od po tri varijable. Prva grupa varijabli se koristi da se formira matrica **M** u kojoj svaka vrednost pokazuje udaljenost između jedinica posmatranja. Na primer, m_{12} pokazuje udaljenost između 1. i 2. jedinice. Logično je da je u pitanju simetrična matrica jer m_{21} pokazuje istu udaljenost između 1. i 2. jedinice. Dijagonalni elementi su nule jer pokazuju udaljenost jedinice posmatranja od sebe same.

Druga grupa varijabli je omogućila formiranje matrice udaljenosti **E**, koja ima iste karakteristike kao i prethodna matrica u pogledu simetričnosti i dijagonalnih elemenata.

Mantel-ov test ocenjuje da li su elementi dve matrice u značajnoj korelaciji. Statistika testa koje se ponekad koristi je koeficijent korelacije između odgovarajućih elemenata (m_{11} sa e_{11}, m_{12} sa e_{12} itd.). Uzimaju se zapravo samo elementi ispod glavne dijagonale. U nxn matrici ima ukupno $n(n\text{-}1)/2$ elemenata ispod glavne dijagonale.

Druga statistika testa je sledeća:

Ova statistika se upoređuje sa distribucijom Z koja se dobija formiranjem niza na slučajan način od elemenata jedne od matrica. Zbog toga se i kaže da je ovo test randomizacije.