ПУШКИЦЕ
ФОН андерграунд

LINEARNI STATISTIČKI MODELI

SKRIPTA – PITANJA ZA PISMENI DEO ISPITA

Jun 2018. godine

SPISAK ISPITNIH PITANJA – JUN 2018.

1. Metode zavisnosti
2. Metode međusobne zavisnosti
3. Vrste podataka i merne skale
4. Kovarijaciona i korelaciona matrica slučajnog vektora X
5. Diskriminaciona analiza – osnovna ideja i ciljevi
6. Metod glavnih komponenata – osnovna ideja i ciljevi
7. Definicija i osobine glavnih komponenata
8. Izbor broja glavnih komponenata
9. Faktorska analiza – osnovna ideja i ciljevi
10. Model faktorske analize
11. Određivanje broja faktora
12. Rotacija faktora
13. Interpretacija faktora
14. Analiza grupisanja – osnovna ideja i ciljevi
15. Hijerarhijski i nehijerarhijski metodi grupisanja
16. Testiranje nezavisnosti kategorijskih obeležja (Hi-kvadrat test nezavisnosti)
17. Testiranje nezavisnosti kvantitativnih obeležja (testiranje koeficijenta korelacije)
18. T-test nezavisnih uzoraka
19. Man-Vitnijev test
20. Analiza varijanse

Sadržaj pitanja je u najvećoj meri preuzet iz knjige „Multivarijaciona analiza“ – Zlatka Kovačića, kao i određenih materijala sa Matematičkog fakulteta. Materijal je namenjen pripremi za pismeni ispit iz predmeta Linearni statistički modeli.

1. Metode zavisnosti

Klasifikacije metoda multivarijacione analize zasnovane su na različitim klasifikacionim kriterijumima. Kod jednog od pristupa, u cilju poređenja dva objekta, posmatramo redove u matrici podataka, odnosno definišemo različite mere bliskosti između dva objekta ili osobe. Osnovu ovih metoda multivarijacione analize predstavlja **matrica odstojanja** između objekata.

Metode se prema ovom pristupu dele u dve grupe: metode zavisnosti i metode međuzavisnosti. Ukoliko se u istraživanju bavimo ispitivanjem zavisnosti između dva skupa promenljivih, gde jedan skup predstavlja zavisne promenljive, a drugi nezavisne, tada je reč o **metodama zavisnosti**.

U metode zavisnosti spadaju:

1. **Multivarijaciona regresija –** razlikujemo dva slučaja: prvi, kada se bavimo analizom zavisnosti jedne promenljive (zavisne promenljive) od skupa drugih (nezavisnih promenljivih) – ovaj slučaj predstavlja **metod višestruke regresije.** Drugi slučaj je kada skup zavisnih promenljivih sadrži više od jednog člana, te tada govorimo o opštijem modelu multivarijacione regresije. Zadatak je oceniti ili predvideti srednju vrednost zavisnih promenljivih, na osnovu poznatih vrednosti nezavisnih.
2. **Kanonička korelaciona analiza –** njome se uspostavlja linearna zavisnost između skupa nezavisnih i skupa zavisnih promenljivih. Kod izračunavanja kanoničke korelacije, formiraju se dve linearne kombinacije, po jedna za skup nezavisnih i zavisnih promenljivih, a koeficijent korelacije između njih treba da bude maksimalan.
3. **Diskriminaciona analiza –** bavi se razdvajanjem grupa i alokacijom opservacija u ranije definisane grupe. Ona omogućava da otkrijemo koja je promenljiva doprinela najviše da se razdvoje grupe, kao i da predvidi verovatnoću da će neki objekat pripasti nekoj od grupa.
4. **Multivarijaciona analiza varijanse (MANOVA) –** koristi se kada nam je cilj da ispitamo uticaj različitih nivoa jedne ili više „eksperimentalnih“ promenljivih na dve ili više zavisnih promenljivih. Koristi nam u situaciji kada je moguće sprovesti kontrolisani eksperiment. Osnovni cilj je testiranje hipoteze koja se tiče varijanse efekata dve ili više zavisnih promenljivih.
5. **Logit analiza –** kada je u regresionom modelu zavisna promenljiva dihotomnog tipa (npr. pol može da bude muški i ženski), tada takav model predstavlja regresioni model sa kvalitativnom zavisnom promenljivom. Kod njih je zavisna promenljiva zapravo **logit funkcija** – logaritam količnika verovatnoća da će dihotomna zavisna promenljiva uzeti jednu ili drugu vrednost. Ove modele nazivamo i modelima logističke regresione analize.

2. Metode međusovne zavisnosti

Ako nema osnova za podelu promenljivih na dva skupa (zavisne i nezavisne), tada se koriste **metode međuzavisnosti**. U ove metode spadaju:

1. **Analiza glavnih komponenti –** služi za redukciju većeg broja promenljivih koje posmatramo na manji broj novih promenljivih koje nazivamo glavne komponente. Najčešće uz pomoć manjeg broja glavnih komponenata objašnjavamo najveći deo varijanse originalnih promenljivih, što omogućava lakše razumevanje podataka. Zadatak je napraviti linearnu kombinaciju originalnih promenljivih (glavnih komponenata) uz uslov da one treba da obuhvate što je moguće veći iznos varijanse početnog skupa promenljivih.
2. **Faktorska analiza –** slična je metodi glavnih komponenti, jer koristi opis varijacija između promenljivih na osnovu manjeg broja promenljivih (nazivamo ih **faktorima**). Međutim, za razliku od prethodne metode, faktorska analiza pretpostavlja postojanje odgovarajućeg statističkog modela kojim originalnu promenljivu iskazujemo kao linearnu kombinaciju faktora plus grešaka modela. Na ovaj način se celokupna kovarijansa ili korelacija objašnjava zajedničkim faktorima, a neobjašnjeni deo se pridružuje grešci – **specifičnom faktoru**. Ovde težimo da objasnimo kovarijansu, odnosno onaj deo ukupne varijanse koji promenljiva deli sa ostalim promenljivama.
3. **Analiza grupisanja –** služi za redukciju podataka, ali za razliku od prethodne dve metode, ona je orijentisana ka redovima matrice podataka (objektima). Ovom analizom kombinujemo objekte u grupe relativno homogenih objekata, a zadatak je identifikovanje manjeg broja grupa, tako da elementi koji pripadaju nekoj grupi budu što sličniji jedan drugom.
4. **Višedimenziono proporcionalno prikazivanje –** orijentisano je ka objektima, a koristi meru sličnosti, odnosno razlike između njih, u cilju njihovog prostornog prikazivanja. Prostorna reprezentacija sadrži geometrijski raspored tačaka na mapi, gde se svaka tačka odnosi na jedan od objekata. Ukoliko smo za računanje mere sličnosti koristili kvanitativne promenljive, metodi dodajemo pridev **kvantitativna**, a ako smo koristili kvalitativne, onda dodajemo pridev **kvalitativna**.
5. **Loglinearni modeli –** omogućuju ispitivanje međusobne zavisnosti kvalitativnih promenljivih koje formiraju višedimenzionu tabelu kontigencije. Ukoliko je jedna od promenljivih u tabeli zavisna, onda na osnovu ocenjenih loglinearnih modela možemo izvesti logit modele. Ali, ovde se logit funkcija izražava preko ćelijskih frekvencija, za razliku od logit modela.

3. Vrste podataka i merne skale

Statistička obeležja mogu biti kvantitativna (merljiva) ili kvalitativna (nemerljiva). Kvantitativne promenljive su one kod kojih se vrednosti razlikuju po veličini, a kvalitativne promenljive su one kod kojih se vrednosti razlikuju po vrsti. Klasifikaciju metoda multivarijacione analize moguće je izvršiti i prema vrsti podataka koji se koriste.

		Kvantitativne promenljive	Kvalitativne promenljive
Metode međuzavisnosti		Glavne komponente Faktorska analiza Analiza grupisanja Kvantitativno višedimenziono proporcionalno prikazivanje	Loglinearni modeli Kvalitativno višedimenziono proporcionalno prikazivanje
Metode zavisnosti	Jedna zavisna promenljiva	Višestruka korelacija Višestruka regresija	Diskriminaciona analiza (zavisna promenljiva je kvalitativna) Logit analiza
	Više zavisnih promenljivih	Višedimenziona regresija Višedimenziona analiza varijanse Kanonična korelaciona analiza	Kanonična korelaciona analiza sa veštačkim promenljivim

Merenja kvantitativnog obeležja iskazujemo na različitim skalama i u različitim jedinicama mere. Ukoliko se jedinica mere može beskonačno deliti (primer: kilometri, metri, centimetri), tada kažemo da je promenljiva **neprekidna**. Kada jedinica mere nije deljiva (primer: veličina porodice), tada promenljivu nazivamo **prekidnom.** Najčešće korišćena skala kod kvantitativnih promenljivih je **skala odnosa.** Ona ima sledeće osobine: količnik ma koje dve vrednosti ima smislenu interpretaciju, rastojanje između dva objekta mereno na ma kom delu ove skale je jednako i opservacijama pozicioniranim na ovoj skali mogu se dodeliti rangovi od višeg ka nižim.

Postoji i **intervalna skala**, koja nema fiksni početak. Temperaturna skala je primer intervalne skale, a za nju važe samo poslednje dve osobine koje važe za skalu odnosa. Kod kvantitativnih obeležja poslednji tip skale je **ordinalna skala**, za koju važi samo poslednja osobina.

Najniži nivo merne skale koriste kvalitativna obeležja i naziva se **nominalna skala.** Ona ne omogućava ni rangiranje jedinica. Kod nje kategorijama pridružujemo vrednosti, kako bismo ih kodirali radi lakše obrade. Primer: bračni status može imati kategorije: neoženjen, oženjen, razveden, udovac, razdvojen – pridružujemo im vrednosti 1, 2, 3, 4 i 5.

4. Kovarijaciona i korelaciona matrica slučajnog vektora X

Za ma koji par slučajnih promenljivih X_j i X_k definišemo kovarijansu: $\sigma_{jk} = E[(X_j - \mu_j)(X_k - \mu_k)]$ – nju označavamo i kao $Cov(X_j, X_k)$, pri čemu je na osnovu definicije: $Cov(X_j, X_j) = Var(X_j)$ i $Cov(X_j, X_k) = Cov(X_k, X_j) = \sigma_{kj} = \sigma_{jk}$.

Za slučajan vektor **X** definišemo (p x p) simetričnu matricu kod koje je j-ti dijagonalni element $\sigma_{jj} = Var(X_j)$, a čiji je (j,k) – element $\sigma_{jk} = Cov(X_j, X_k)$, $j \neq k$. Ovu matricu nazivamo **kovarijacionom matricom** od **X** i označavamo je kao Var(**X**) ili Cov(**X**), odnosno **∑.** Tako je:

$$Cov(\mathbf{X}) = \boldsymbol{\Sigma} = [\sigma_{jk}] = \left(\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}\right) = \left(\begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_p) & Cov(X_2, X_p) & \cdots & Var(X_p) \end{bmatrix}\right)$$

Kovarijacionu matricu možemo iskazati i kao očekivanu vrednost slučajne matrice. Za slučajan vektor **X** sa sredinom **μ** definišemo (p x p) simetričnu slučajnu matricu kvadrata, odnosno uzajamnih proizvoda odstupanja elemenata slučajnog vektora od odgovarajuće sredine. Slučajna matrica je proizvod slučajnih vektora odstupanja od sredine, tj. **(X - μ)(X - μ)'**, pa je njena očekivana vrednost:

$$E[\mathbf{(X - \mu)(X - \mu)'}] = \boldsymbol{\Sigma}$$

Koeficijent korelacije između dve slučajne promenljive X_j i X_k definišemo kao:

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}}$$

što predstavlja normalizovanu kovarijansu između X_j i X_k. On uzima vrednost iz intervala -1 do +1. Ukoliko koeficijent korelacije uzme donju ili gornju graničnu vrednost, tada kažemo da postoji perfektrna linearna veza između X_j i X_k.

Korelacionu matricu ***ρ*** možemo dobiti na osnovu poznate kovarijacione matrice, a njen (j,k) – ti element definisan je gornjim izrazom. U matričnoj notaciji veza između korelacione i kovarijacione matrice je data sa:

$$\boldsymbol{\rho} = (\mathbf{D}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{D}^{1/2})^{-1} = \left(\begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}\right)$$

gde smo sa **D** označili dijagonalnu matricu koja sadrži elemente na glavnoj dijagonali kovarijacione matrice **∑**.

5. Diskriminaciona analiza – osnovna ideja i ciljevi

Metod multivarijacione analize koji se bavi razdvajanjem različitih grupa i alokacijom opservacija u unapred definisane grupe nazivamo **diskriminaciona analiza**.

Diskriminaciona analiza ima dva osnovna cilja. Prvi cilj je da se utvrdi da li postoji statistički značajna razlika u sredinama dve ili više grupa, a zatim da odredi koja od promenljivih daje najveći doprinos utvrđenoj razlici. Ovaj cilj analize nazivamo **diskriminacija** ili **razdvajanje** između grupa. Drugi cilj se odnosi se na utvrđivanje postupka za klasifikaciju opservacija na osnovu vrednosti nekoliko promenljivih u dve ili više razdvojenih, unapred definisanih grupa. Ovaj cilj analize nazivamo **klasifikacija** ili **alokacija** opservacija.

Sa tehničke strane, osnovni cilj diskriminacione analize jeste formiranje linearnih kombinacija nezavisnih promenljivih kojima će se diskriminacija između unapred definisanih grupa izvršiti tako da greška pogrešne klasifikacije opservacija bude minimalna. Linearnom kombinacijom nezavisnih promenljivih za svakog ispitanika formiramo broj koji se naziva **diskriminacioni skor**, koji se zatim transformiše u verovatnoću da ispitanik potiče iz jedne od grupa. U opštem slučaju, imamo da je:

$$\mathbf{Y} = \mathbf{a'X}$$

gde je **Y** diskriminacioni skor, dok je **a** p-dimenzioni vektor diskriminacionih koeficijenata (koeficijenti linearne kombinacije), a **X** je p-dimenzioni vektor nezavisnih promenljivih. Projekcija tačaka sa dijagrama rasturanja na y-osu generiše jednodimenzione rasporede diskriminacionih skorova dveju populacija π_1 i π_2. Sredine diskriminacionih skorova za ove dve grupe predstavljaju prvi, odnosno drugi **centroid**. Njihovim međusobnim poređenjem možemo utvrditi koliko su grupe udaljene jedna od druge.

U diskriminacionoj analizi formiramo linearnu kombinaciju merljivih promenljivih, ali je zavisna promenljiva nemerljiva (kvalitativna). Za razliku od regresione analize, ovde je zavisna promenljiva fiksna (uzima vrednosti 0 i 1 ako razmatramo problem diskriminacije dve grupe), a nezavisne promenljive su slučajne promenljive koje su normalno raspoređene.

6. Metoda glavnih komponenata – osnovna ideja i ciljevi

Metod multivarijacione analize koji se koristi za smanjivanje dimenzije skupa podataka (sačinjava ga veliki broj uzajamno korelisanih promenljivih) uz istovremeno zadržavanje maksimalno mogućeg varijabiliteta koji je prisutan u tim podacima, naziva se **metod glavnih komponenata.** Kažemo da ovaj metod pored toga što redukuje dimenziju skupa podataka predstavlja i istraživačko sredstvo analize pomoću koga se generišu hipoteze o proučavanom fenomenu.

Osnovni zadatak metode glavnih komponenata jeste određivanje one linearne kombinacije originalnih promenljivih koja će imati maksimalnu varijansu. Drugi, opštiji zadatak ove metode jeste određivanje nekoliko linearnih kombinacija originalnih promenljivih koje će, pored toga što imaju maksimalnu varijansu, biti među sobom nekorelisane, gubeći u što je manje mogućoj meri informaciju sadržanu u skupu originalnih promenljivih.

U ovom postupku, originalne promenljive se transformišu u nove promenljive koje nazivamo **glavne komponente.** *Prva glavna komponenta* je konstruisana tako da obuhvata najveći deo varijanse, a naredne onaj deo koji nije još uvek obuhvaćen.

Ovime se postižu dva cilja:

1. Vrši se redukcija originalnog skupa podataka
2. Olakšava se njihova interpretacija

Problem se može prikazati i grafički, koristeći proizvoljne linearne kombinacije.

Ako se zahteva reprezentovanje dvodimenzionalnog skupa samo jednom promenljivom, onda bismo izabrali onu koja ima veći varijabilitet. Na osnovu promenljive sa većim varijabilitetom možemo u većoj meri razlikovati pojedinačne opservacije dvodimenzionog skupa. U ekstremnom slučaju kada sve tačke leže na pravoj normalnoj na X_1 osu, tada je dovoljno analizirati samo promenljivu X_2, jer ona nosi svu informaciju o varijabilitetu dvodimenzionog skupa podataka. Naš izbor koeficijenata se može opisati kao zadatak maksimiziranja varijanse linearne kombinacije uz uslov da je zbir kvadrata koeficijenata linearne kombinacije jednak jedinici. Geometrijski to znači da je vektor koeficijenata linearne kombinacije $\mathbf{[\alpha_{11}, \alpha_{12}]'}$ jedinične dužine. Izborom koeficijenata, mi zapravo menjamo ugao pod kojim se projektuju tačke na pravu liniju.

7. Definicija i osobine glavnih komponenata

Pretpostavimo da je **X** p – dimenzioni slučajan vektor sa kovarijacionom matricom **∑.** Neka je $Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + ... + \alpha_{1p}X_p = \boldsymbol{\alpha'_1 X}$ linearna kombinacija elemenata slučajnog vektora **X**, gde su $\alpha_{11}, \alpha_{12}, ... , \alpha_{1p}$ koeficijenti linearne kombinacije. Poznato je da je $Var(Y_1) = \boldsymbol{\alpha'_1 \sum \alpha_1}$, pa je naš zadatak da odredimo vektor koeficijenata $\boldsymbol{\alpha_1}$ tako da se maksimizira varijansa od Y_1. Pri tome vodimo računa o ograničenju da je vektor koeficijenata jedinične dužine – $\boldsymbol{\alpha'_1 \alpha_1} = 1$.

Problem se rešava pomoću maksimizacije Lagranžove funkcije:

$$\boldsymbol{\alpha'_1 \sum \alpha_1} - \lambda(\boldsymbol{\alpha'_1 \alpha_1} - 1)$$

gde je λ Lagranžov činilac. Diferenciranjem Lagranžove funkcije po koeficijentima $\boldsymbol{\alpha_1}$, a zatim izjednačavanjem dobijenog izraza sa nulom, dobijamo:

$$\boldsymbol{\sum \alpha_1} - \lambda \boldsymbol{\alpha_1} = 0$$

ili

$$(\boldsymbol{\sum} - \lambda \mathbf{I})\boldsymbol{\alpha_1} = 0$$

gde je **I** (p x p) jedinična matrica. Da bi se dobilo netrivijalno rešenje za $\boldsymbol{\alpha_1}$ determinanta $|\boldsymbol{\sum} - \lambda \mathbf{I}|$ mora biti jednaka nuli. Pošto težimo da maksimiziramo varijansu, za λ ćemo uzeti najveći karakteristični koren, a njemu je pridružen odgovarajući vektor $\boldsymbol{\alpha_1}$. Ako nam je zadatak da odredimo više od jedne linearne kombinacije, tada postupamo kao i u ovom slučaju, pri čemu uzimamo dodatni uslov da kovarijansa prve i druge glavne komponente bude jednaka nuli.

Glavne komponente imaju sledeće osobine:

$$E(Y_j) = 0,\ Var(Y_j) = \lambda_j,\ Cov(Y_i, Y_j) = 0,\ i \neq j$$

$$Var(Y_1) \geq Var(Y_2) \geq ... \geq Var(Y_p) \geq 0$$

Takođe, može se dokazati da su generalizovane varijanse glavnih komponenata jednake generalizovanim varijansama originalnog skupa promenljivih. Neka je, dakle, **Y** vektor glavnih komponenata takav da je $\mathbf{Y'} = [Y_1, Y_2, ... , Y_p]$. Sada se transformacija originalnog skupa promenljivih sadržanog u vektoru **X** može pisati na sledeći način: **Y** = **AX**, gde je **A** matrica čiji su redovi karakteristični vektori kovarijacione matrice **∑**. Matrica **A** ima osobinu da je $\mathbf{A'} = \mathbf{A^{-1}}$, pa se **Y** = **AX** naziva **ortogonalna transformacija** ili **rotacija**, a sama matrica **A** se naziva ortogonalnom matricom, a njena osobina je i da je $|\mathbf{A}| = \pm 1$. Transformacija se naziva ortogonalnom, jer se njome vrši rotacija koordinatnih osa za određen ugao, pri čemu ose ostaju normalne jedna na drugu, a ugao između bilo koja dva vektora nakon rotacije ostaje isti.

8. Izbor broja glavnih komponenata

Jedan od ciljeva analize glavnih komponenata jeste redukcija početnog skupa podataka. Umesto velikog broja promenljivih u daljoj analizi se koristi samo manji broj glavnih komponenata koje u najvećoj meri treba da obuhvate varijansu početnog skupa podataka. Postoje brojni pristupi koji se tiču odabira odgovarajućeg broja glavnih komponenata.

Prvi pristup polazi od fiksiranja kumulativne proporcije ukupne varijanse koja je „objašnjena" izdvojenim skupom glavnih komponenata. Obično se izabere proporcija od 80% ili 90% ukupne varijanse, pa se broj zadržanih glavnih komponenata povećava dok se ne postigne ovaj kriterijum.

Drugi pristup sugeriše da treba zadržati one glavne komponente čija varijansa (λ_j) je veća od prosečne vrednosti

$$\bar{\lambda} = \sum_{j=1}^{p} \frac{\lambda_j}{p}.$$

Ako umesto kovarijacione koristimo korelacionu matricu, tada je prosečna vrednost varijanse jednaka jedinici, što znači da bi kriterijum glasio: treba zadržati one glavne komponente kod kojih je varijansa veća od jedinice.

Prema trećem pristupu u kriterijumu izbora neophodno je koristiti geometrijsku sredinu. Generalizovana varijansa je jednaka proizvodu karakterističnih korena, odnosno

$$\prod_{j=1}^{p} \lambda_j.$$

Ako dobijenu vrednost dignemo na stepen 1 / p dobićemo geometrijsku sredinu karakterističnih korena. Prosečna generalizovana varijansa data je geometrijskom sredinom karakterističnih korena, pa zadržavamo one komponente čiji je karakteristični koren veći od geometrijske sredine svih karakterističnih korena.

Poslednji pristup se zasniva na grafičkom prikazu vrednosti karakterističnih korena prema njihovom rednom broju. Ovaj dijagram se naziva „**scree test**" – prelom na krivoj se određuje tako što se lenjir prisloni uz poslednje vrednosti karakterističnog korena proveravajući da li one leže na pravoj liniji. Broj glavnih komponenata određujemo tako što uočavamo tačku nakon koje spomenuta prava ima prelom, pri čemu se krećemo od većeg ka manjem rednom broju glavne komponente. Broj glavnih komponenata zapravo predstavlja redni broj glavne komponente čija vrednost karakterističnog korena kao poslednja leži na pravoj liniji.

9. Faktorska analiza – osnovna ideja i ciljevi

Metod multivarijacione analize koji se koristi za opisivanje međusobne zavisnosti velikog broja promenljivih korišćenjem manjeg broja osnovnih, ali neopažljivih slučajnih promenljivih poznatih kao **faktori** naziva se **faktorska analiza**. U faktorskoj analizi nas zanimaju vandijagonalni elementi (kovarijanse), za razliku od analize glavnih komponenata. Faktorska analiza podrazumeva postojanje teorijskog modela kojim se uspostavlja relacija između opservacija dimenzione promenljive i manjeg broja zajedničkih faktora.

Osnovna ideja faktorske analize sastoji se u sledećem – ona je razvijena kako bi se lakše analizirali rezultati određenih testova. Ako za primer uzmemo test inteligencije, zadatak faktorske analize je da utvrdi da li se inteligencija sastoji iz jednog opšteg faktora ili od nekoliko zajedničkih faktora.

Rezultate svih testova (X_i) moguće je iskazati u obliku:

$$X_i = \beta_i F + \varepsilon_i,\ i = 1, 2, 3, \ldots, n$$

U ovom modelu F je zajednički faktor, a β_i su koeficijenti koje nazivamo **faktorska opterećenja**, a ε_i su slučajne greške, odnosno **specifični faktori.** Rezultati ovih testova mogu se dekomponovati na dva dela, pri čemu se jedan odnosi na sve testove (F), a drugi je specifičan za svaki test (ε_i).

Kasnija istraživanja su proširila prvobitni model, te je uvedeno nekoliko zajedničkih faktora, a specifičan faktor je razložen na dva dela. Dakle, faktorska analiza služi za redukciju originalnog skupa podataka – koristimo je da bismo identifikovali zajedničku strukturu koja je generisala dobijeni skup korelisanih originalnih promenljivih. To je **istraživačka primena** faktorske analize, ona se koristi u deskriptivne svrhe. Druga primena se tiče istraživanja gde posedujemo već neku teorijsku informaciju o zajedničkoj strukturi, a faktorsku analizu koristimo kako bismo testirali hipoteze o broju zajedničkih faktora. Dakle, ona se ovde koristi kako bi se potvrdila, odnosno negirala hipoteska struktura podataka.

Za razliku od analize glavnih komponenata, faktorska analiza polazi od razlagaja promenljive na dva dela: zajednički i specifični. Zajednički deo je onaj deo varijacija promenljive koji ona deli sa ostalim promenljivama, dok je specifilan onaj deo varijacija koji je poseban za tu promenljivu. Faktorska analiza izučava deo varijacija koji je zajednički za sve, a analiza glavnih komponenata ukupan varijabilitet.

10. Model faktorske analize

Pretpostavimo da je **X** p – dimenzioni vektor opažljivih promenljivih sa sredinom **μ** i kovarijacionom matricom **∑**. Model faktorske analize pretpostavlja da se **X,** vektor opažljivih promenljivih, može izraziti preko skupa od m neopažljivih promenljivih, koje nazivamo **zajednički faktori**, u oznaci F_1, F_2, ... ,F_m, gde je m << p i p specifičnih, ali neopažljivih faktora, u oznaci ε_1, ε_2, ... , ε_p. Model u razvijenog obliku dat je sledećim jednačinama:

$$(X_1 - \mu_1) = \beta_{11}F_1 + \beta_{12}F_2 + ... + \beta_{1m}F_m + \varepsilon_1$$

$$(X_2 - \mu_2) = \beta_{21}F_1 + \beta_{22}F_2 + ... + \beta_{2m}F_m + \varepsilon_2$$

$$(X_p - \mu_p) = \beta_{p1}F_1 + \beta_{p2}F_2 + ... + \beta_{pm}F_m + \varepsilon_p$$

ili ekvivalentno u matričnoj notaciji:

$$\mathbf{X - \mu = \quad B \quad F \quad + \quad \varepsilon}$$

(px1) (pxm) (mx1) (px1)

gde je:

$$\boldsymbol{X - \mu} = \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix}, \boldsymbol{F} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & ... & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pm} \end{bmatrix}$$

Elementi matrice B nazivaju se **faktorska opterećenja** i-te promenljive na j-ti faktor, a sama matrica se naziva **matrica faktorskih opterećenja**. Na prvi pogled, model faktorske analize više liči na model višestruke regresije. Međutim, ovde p odstupanja $(X_1 - \mu_1)$, ... , $(X_p - \mu_p)$ izražavamo preko m + p slučajnih promenljivih F_1, F_2, ... , F_m i ε_1, ε_2, ... , ε_p koje su neopažljive, za razliku od regresionog modela gde su nezavisne promenljive opažljive.

Dodajemo dodatna ograničenja, vezana za zajedničke faktore:

$$E(\mathbf{F}) = 0, Cov(\mathbf{F}) = E(\mathbf{FF'}) = \mathbf{\Phi}$$

Što se specifičnih faktora tiče, njihova očekivana vrednost je jednaka nuli, a kovarijaciona matrica dijagonalna:

$$E(\boldsymbol{\varepsilon}) = 0, Cov\,(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon\varepsilon'}) = \Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi_p \end{bmatrix}$$

Takođe, pretpostavlja se da su zajednički faktori nezavisni od specifičnih, odnosno da je:

$$Cov(\boldsymbol{\varepsilon, F}) = E(\boldsymbol{\varepsilon F'}) = 0$$

Vezu između odstupanja opažljivih promenljivih od njihove sredine i neopažljivih faktora zajedno sa navedenim pretpostavkama i ograničenjima nazivamo **model faktorske analize**. Ovaj model omogućava razlaganje kovarijacione matrice **∑** na:

$$\boldsymbol{\Sigma} = \mathbf{BB'} + \boldsymbol{\Psi}$$

Korelacionu matricu promenljivih **X** i faktora **F** nazivamo **matricom faktorske strukture.**

Na osnovu razlaganja kovarijacione matrice, imamo da je varijansa i – te promenljive:

$$Var(X_i) = \sigma_{ii} = \beta_{i1}^2 + \beta_{i2}^2 + ... + \beta_{im}^2 + \Psi_i = \sum_{j=1}^{m} \beta_{ij}^2 + \Psi_i \text{ , } i = 1, 2, ... , p$$

Znači da je varijansa i-te originalne promenljive podeljena na dva dela. Prvi deo je varijansa objašnjena zajedničkim faktorima i nazivamo ga **zajednička varijansa** ili **komunalitet** (u oznaci h_i^2), a drugi deo nazivamo **specifična varijansa** (u oznaci Ψ_i).

Imamo da je jedinična varijansa standardizovane promenljive jednaka:

$$Var(X_i) = 1 = h_i^2 + \Psi_i$$

Takođe, generalizovana varijansa od **X** je:

$$tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} = + \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_{ij}^2 + \sum_{i=1}^{p} \psi_i$$

Ako sa **h** označimo ukupan komunalitet od **X**, tada je:

$$tr(\boldsymbol{\Sigma}) = h + tr(\boldsymbol{\Psi})$$

što znači da je ukupna ili generalizovana varijansa od **X** jednaka zbiru ukupnog komunaliteta i ukupne varijanse specifičnih faktora.

11. Određivanje broja faktora

Iako se faktorska analiza razlikuje od metode glavnih komponenata, postupci izbora broja glavnih komponenata se koriste i prilikom određivanja broja faktora. Najpoznatiji je kriterijum jediničnog korena gde zadržavamo u modelu onoliko zajedničkih faktora koliko ima karakterističnih korena uzoračke korelacione matrice koji su veći od jedinice. Ovo se koristi kada je broj promenljivih između 20 i 50, ali ako je broj promenljivih veći od 50, tada ovaj kritetijum bira previše zajedničkih faktora, a ako je broj ispod 20, tada bira mnogo mali broj zajedničkih faktora.

Za određivanje broja faktora može se koristiti i „scree test“ koji izdvaja veći broj faktora nego prethodni metod. Preporuka je da se koristi više od jednog metoda za odabir broja faktora.

Moguće je koristiti i određene statističke testove za određivanje broja odgovarajućih faktora, pri čemu za veliki broj zajedničkih faktora statistika testa nije pouzdana, pa se zato predlaže da se koristi postupak korak po korak, tako što će analiza započeti sa jednim zajedničkim faktorom, pa se potom broj faktora povećava za po jedan sve dok se ne prihvati nulta hipoteza, koja glasi: H_0: $\mathbf{BB' + \Psi}$ (alternativna hipoteza je: H_1: $\sum$) ili dok broj stepena slobode ne postane negativan. Mora se ipak proveriti da li je uopšte potrebno sprovesti faktorsku analizu, pošto ukoliko je kovarijaciona matrica dijagonalna, to znači da su originalne promenljive međusobno nekorelisane, pa nema potrebe za faktorskom analizom. Zato se prvo proverava da li se može odbaciti hipoteza o sferičnosti.

12. Rotacija faktora

U faktorskoj analizi ortogonalnu transformaciju matrice faktorskih opterećenja i time impliciranu ortogonalnu transformaciju faktora (faktorskih osa) nazivamo **rotacija faktora** ili preciznije **ortogonalna rotacija faktora.** Napuštanjem zahteva da rotirani faktori moraju međusobno biti ortogonalni, razvijeni su postupci tzv. **neortogonalne rotacije faktora.** Postupak se primenjuje u cilju dobijanja takve matrice faktorskih opterećenja koja će olakšati interpretaciju faktora. Izbor ugla za koji ćemo rotirati faktore opredeljen je jednim od kriterijuma, a najčešće se koristi onaj pod nazivom **jednostavna struktura.**

Kod jednostavne strukture pokušavamo da postignemo mali broj visokih vrednosti faktorskih opterećenja i veliki broj niskih vrednosti faktorskih opterećenja. Istraživač potom interpretira niske vrednosti kao nule, a visoke kao vrednosti različite od nule.

Ortogonalna rotacija faktora ne menja međusobni odnos faktorskih osa, one su i dalje ortogonalne. Ona se po tome razlikuje od neortogonalne rotacije faktora kod koje nema tog ograničenja, jer se faktorske ose rotiraju nezavisno jedna od druge.

Neka je **T** ortogonalna matrica kojom smo transformisali ocenjenu matricu faktorskih opterećenja $\widehat{\boldsymbol{B}}$**.** Znači da je $\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{B}}\boldsymbol{T}$, pri čemu je **T'T** = **TT'** = **I**, gde matricu $\widehat{\boldsymbol{\Gamma}}$ nazivamo ocenjena matrica rotiranih faktorskih opterećenja.

Najčešće korišćenjen analitički metod ortogonalne rotacije faktora je Kaiserov **varimax metod**. Kod njega posmatramo kvadrate elemenata matrice $\widehat{\boldsymbol{\Gamma}}$ u j-oj koloni. Sabirajući vrednosti varijansi kod svih m faktora dobijamo **sirov varimax kriterijum**, a na osnovu njega dobijamo **normalan varimax kriterijum.** Postupak primene varimax i drugih kriterijuma jednostavnosti strukture je iterativan proces. Izdvojeni faktori se posmatraju po parovima i vrši se njihova rotacija dok se ne postigne maksimalna vrednost varimax kriterijuma za prvi par faktora. Zatim se prvi rotirani faktor u paru sa trećim, nerotiranim faktorom, rotira do postizanja maksimuma svih varimax kriterijuma. Postupak se ponavlja sve dok se svih m(m-1)/2 parova faktora na navedeni način rotiraju. Ovaj niz rotacija se naziva **ciklus**. On se ponavlja sve dok se ne postigne da su svi uglovi dobijeni za parove faktora manji od unapred izabrane vrednosti, koja predstavlja kriterijum konvergencije.

Za rotaciju se koriste i druge metode, kao što je **quatrimax kriterijum** prema kome se kao indikator jednostavnosti strukture uzima suma varijansi kvadrata svih elemenata matrice $\widehat{\boldsymbol{\Gamma}}$**.** Ova metoda obično rezultira u opštem faktoru, jer se varijansa računa na osnovu svih elemenata matrice faktorskih opterećenja.

Postoji i tzv. **ortomax metod** koji se zasniva na ponderisanom proseku sirovog varimax i quatrimax kriterijuma i on predstavlja generalizaciju ortogonalnih kriterijuma rotacije. Posebni slučajevi ovog kriterijuma su **biquatrimax** i **equamax** kriterijum.

13. Interpretacija faktora

Pre nego što se pristupi interpretaciji faktora mora se odgovoriti na sledeće pitanje: koji se od ocenjenih elemenata matrice faktorskih opterećenja mogu smatrati statistički značajnim? Na raspolaganju je nekoliko iskustvenih kriterijuma.

Prvi je proistekao iz iskustva velikog broja istraživača u primeni modela faktorske analize. Oni su sugerisali da se svi koeficijenti faktorskih opterećenja čija je apsolutna vrednost veća od 0.30, smatraju statistički značajno različitim od nule. Kod veličine uzorka od 50 i više elemenata, ovaj kriterijum se pokazao prihvatljivim.

Drugi kriterijum je zasnovan na činjenici da je kod ortogonalnog modela faktorske analize, matrica faktorskih opterećenja identična matrici faktorske strukture. Kako su elementi ove druge matrice koeficijenti korelacije promenljivih sa faktorima, tako nam njihova visoka vrednost govori da odnosna promenljiva opredeljuje faktor sa kojim je korelisana. Zato se testovi statističke značajnosti koeficijenata korelacije direktno primenjuju na elemente matrice faktorskih opterećenja. Tako, na primer, za t-test za testiranje hipoteze o nultoj vrednosti koeficijenta korelacije sugeriše, za uzorke veličine 100 elemenata i na nivou značajnosti od 5% i 1%, da se smatraju statistički značajnim ona faktorska opterećenja čija je apsolutna vrednost veća od 0.19 i 0.26 respektivno.

Navedeni kriterijumi u obzir ne uzimaju broj promenljivih u analizi, kao i redosled faktora čija opterećenja preispitujemo sa stanovišta značajnosti.

Postupak interpretacije faktora je sledeći: posmatramo matricu faktorskih opterećenja po redovima, zaokružujemo koeficijente sa najvećom apsolutnom vrednošću u prvom redu, pa prelazimo u red ispod i tako postupamo sa svim preostalim redovima matrice. Nakon toga proveravamo značajnost zaokruženih faktorskih opterećenja korišćenjem nekog od prethodno navedenih kriterijuma, pa podvučemo statistički značajna faktorska opterećenja. Idealna situacija je kada se broj zaokruženih i podvučenih koeficijenata poklapa, jer tada svaka promenljiva pripada samo jednom faktoru. Svakom faktoru potom pridružujemo naziv, s obzirom na strukturu faktora, tj. listu promenljivih koje su visoko korelisane sa tim faktorom.

Moguća su još dva slučaja, da postoji manji broj podvučenih od zaokruženih koeficijenata. To znači da se neka od promenljivih nije pridružila jednom od izdvojenih faktora. Tada možemo zanemariti datu promenljivu ili preispitati njen značaj koristeći njen komunalitet.

Druga situacija je da postoji veći broj statistički značajnih faktorskih opterećenja u jednom redu. To znači da je promenljiva korelisana sa više faktora, što otežava interpretaciju. To se najćešće dešava kada imamo nerotiranu matricu faktorskih opterećenja.

14. Analiza grupisanja – osnovna ideja i ciljevi

Metod multivarijacione analize koji se koristi za grupisanje objekata u grupe, tako da su objekti unutar grupe sličniji međusobno, a između grupa znatno različiti, naziva se **analiza grupisanja.**

Da bi odgovorila ovom zadatku, analiza grupisanja zahteva definisanje **mere bliskosti** dva objekta na osnovu njihovih karakteristika. Osnovni zadatak analize grupisanja je nalaženje prirodnog grupisanja skupa objekata ili osboa. Grupisanje objekata je zasnovano na različitim karakteristikama koje merimo kod svakog objekta. Ako smo merili dve karakteristike kod svakog objekta, možemo se poslužiti dijagramom rasturanja u cilju određivanja grupa objekata. Na njemu su objekti unutar grupe slični međusobno (tačke su bliže jedna drugoj), a objekti u različitim grupama različiti (tačke u prostoru su na većoj razdaljini).

Postoji i definicija grupa na osnovu kriterijuma bliskosti, pa se prema njemu smatra da objekti u grupi treba da budu bliži jedni drugima, nego objektima u drugim grupama.

Pored grafičkih metoda, koriste se i analitički postupci na osnovu kojih se prema skupu formalnih pravila vrši grupisanje objekata u grupe. Polaznu osnovu čine podaci uređeni u matricu podataka sa n redova (objekata) i p kolona (promenljivih). Elementi u i-om redu odnose se na različite karakteristike i-og objekta i formiraju njegov **profil**, dok elementi u j-oj koloni predstavljaju vrednosti j-te karakteristike koju različiti objekti uzimaju. Na osnovu ove matrice podataka, formiramo (m x n) **matricu bliskosti** čiji elementi mere stepen sličnosti i razlike između svih parova profila iz matrice podataka. Ona se označava sa **P**, a njeni elementi su p_{rs}, gde je r,s = 1,2, ... , n, a predstavljaju meru bliskosti između r-tog i s-tog objekta.

Nakon formiranja matrice bliskosti, vršimo izbor matrice grupisanja. Metodi grupisanja su skup pravila pridruživanja objekata u grupe na osnovu mere bliskosti između objekata. Najčešće su korišćene hijerarhijske metode grupisanja kod kojih se u svakoj iteraciji objekti pridružuju prethodno formiranim grupama ili sa drugim objektom prave novu grupu. Ovakva struktura predstavlja **hijerarhijsko drvo**. Hijerarhijsku strukturu možemo formirati udruživanjem ili deobom.

Ciljevi analize grupisanja su:

1. **Istraživanje podataka** – otkrivamo strukturu skupa objekata na osnovu analize grupisanja.
2. **Redukcija podataka** – interes je formirati manji broj grupa objekata.
3. **Generisanje hipoteza** – analiza grupisanja nam pomaže da definišemo hipotezi o strukturi podataka.
4. **Predviđanje** – grupe dobijene u analizi grupisanja možemo koristiti u kasnijim istraživanjima u svrhe predviđanja.

15. Hijerarhijski i nehijerarhijski metodi grupisanja

Hijerarhijski metodi

Hijerarhijski metodi grupisanja se mogu svrstati u dve grupe prema tome da li su zasnovani na iterativnom spajanju ili deljenju grupa i objekata. U prvom metodu, od n grupa težimo da napravimo jednu grupu, kod drugog se metod kreće u suprotnom smeru – od jedne grupe koja sadrži sve objekte težimo da iz iste, po određenom kriterijumu, izdvajamo po jedan objekat ili grupu dok se ne formira onoliko grupa koliko ima individualnih objekata.

Najčešće korišćeni metodi grupisanja pripadaju hijerarhijskim metodama udruživanja, a izdvaja se metod povezivanja. Kod **metode jednostrukog povezivanja** polazi se od matrice odstojanja, bira se elemenat koji je najmanji i odgovarajuća dva objekta se udružuju u jednu grupu. Sada se određuje odstojanje nove grupe od ostalih objekata. U drugoj iteraciji, ponovo biramo najmanji element matrice – može se desiti da su neka druga dva objekta bliža međusobno ili da je jedan objekat bliži ranije formiranoj grupi. U prvom slučaju se formira nova grupa, a u drugom se element pridružuje ranije formiranoj grupi. Postupak se nastavlja dok se svi objekti ne udruže u jednu grupu. Ovaj metod povezuje objekte na osnovu najkraćeg odstojanja između njih.

Kod **metode potpunog povezivanja** koraci su identični kao kod metode jednostrukog povezivanja, razlika se javlja jedino u načinu određivanja odstojanja između grupa, jer se kod ove metode odstojanje određuje prema najvećem odstojanju objekata koji pripadaju dvema grupama.

Postoji i tzv. **metod prosečnog povezivanja**, gde su koraci, ponovo identični, a odstojanje se određuje prema prosečnom odstojanju svih objekata koji pripadaju dvema grupama.

Preostala dva metoda hijerarhijskog udruživanja su **metod centroida** i **metod minimalne sume kvadrata** (**Wardov metod**). Kod metoda centroida dve grupe se udružuju u novu ako su njihovi centroidi najmanje udaljeni međusobno u odnosu na međusobnu udaljenost svih mogućih parova grupa koji postoje na datom nivou. Kod Ward-ovog metoda dve grupe se spajaju u jednu, ako je njihovim udruživanjem došlo do najmanjeg povećanja sume kvadrata unutar grupa u odnosu na povećanje sume kvadrata do koga je došlo udruživanjem bilo koje dve druge grupe.

Hijerarhijski metodi se prikazuju pomoću hijerarhijskog drveta, a ako uz njega navedemo i skalu na kojoj su navedene vrednosti mere odstojanja u svakom koraku udruživanja grupa, tada dobijamo **dendogram**. Na osnovu njega možemo formirati **izvedenu matricu odstojanja**. Ovde koristimo i **kofenetički koeficijent** koji je običan koeficijent korelacije između originalnih i izvedenih mera odstojanja.

Nehijerarhijski metodi

Nehijerarhijski metodi dozvoljavaju premeštanje objekata iz ranije formiranih grupa. Do premeštanja će doći ukoliko to sugeriše izabrani kriterijum optimalnosti. U primeni ovih metoda se pretpostavlja da je broj grupa unapred poznati ili ga menjamo tokom postupka grupisanja.

Postupak nehijerarhijskog grupisanja počinje podelom skupa objekata u izabran broj grupa. Alternativna podeli objekata je određivanje inicijalne **klice**, odnosno centroida za svaku grupu. Potom se određuje odstojanje između svakog objekta i grupe. Objekti se razmeštaju u grupe koje su najbliže, nakon pridruživanja se izračunava centroid grupe iz koje je objekat izašao i grupe u kojoj se objekat pridružio. Ponovo se izračunava rastojanje od centroida grupa i vrši se preraspodela objekata, sve dok izabrana funkcija kriterijuma to zahteva. Najpopularniji metod je **metod k-sredina**, prema kome objekat pridružujemo grupi koja ima najbliži centroid.

16. Hi-kvadrat test nezavisnosti

Postupak nazvan hi-kvadrat test se upotrebljava u većini slučajeva ako se radi o kvalitativnim podacima ili ako tim podacima distribucija značajno odstupa od normalne. Već u početku treba naglasiti da se hi-kvadrat test računa samo s frekvencijama, pa u račun nije dopušteno unositi nikakve merne jedinice. Osnovni podaci istraživanja mogu biti i merne vrednosti, ali u hi-kvadrat unose se samo njihove frekvencije.

Hi-kvadrat test je vrlo praktičan test koji može poslužiti onda kad želimo utvrditi da li neke dobijene (opažene) frekvencije odstupaju od frekvencija koje bismo očekivali pod određenom hipotezom. Kod ovog testa nekada tražimo postoji li povezanost između dve varijable i on pokazuje verovatnoću povezanosti. Možemo pretpostaviti da neka teorijska raspodela dobro opisuje opaženu raspodelu frekvencija. Da bismo tu pretpostavku (hipotezu) proverili, primenjujemo ovaj test.

Često želimo znati da li se opažene frekvencije značajno razlikuju od očekivanih frekvencija. Ta razlika se računa se prema sledećoj formuli:

$$\chi^2 = \sum \frac{(f_0 - f_i)^2}{f_i}$$

pri čemu f_0 znači opažene frekvencije, a f_t očekivane (teoretske) frekvencije, tj. frekvencije koje bismo očekivali pod nekom određenom hipotezom. Broj stepeni slobode **ν** definisan je kao broj nezavisnih varijabli uključenih u računanje χ^2.

Nulta hipoteza (H_0) glasila bi: „Opažene frekvencije slede teorijsku raspodelu.“, dok bi alternativna hipoteza (H_1): „Opažene frekvencije ne slede teorijsku raspodelu.“ Nulta hipoteza se odbacuje za ako test značajnosti pokaže da su podaci nekonzistentni sa testom, odnosno za graničnu vrednost testa. Značajnost testa α je verovatnoća odbacivanja nulte hipoteze kada je ona istinita.

17. Testiranje koeficijenta korelacije

Pirsonov r-test je test koji se najčešće koristi kako bi se testirao koeficijent korelacije. Sledeća formula služi za izračunavanje zadatog koeficijenta:

$$r = \frac{N\sum xy \sum(x)(y)}{\sqrt{[N\sum x^2 - \sum(x)^2][N\sum y^2 - \sum(y)^2]}}$$

pri čemu je:

r – Pirosnov (*Pearson*) koeficijent

N – broj opservacija

x, y – zadate promenljive

Ovaj test se može upotrebiti u slučajevima kada želimo da otkrijemo da li postoji statistički značajna razlika između dve promenljive – za primer možemo uzeti vezu između starosti i visine, temperature i prodaje sladoleda, zadovoljstva radnim mestom i zarade i sl.

Obe promenljive treba da imaju normalnu raspodelu.

18. T-test nezavisnih uzoraka

Najčešće upotrebljavan parametarski test značajnosti za testiranje nulte hipoteze je Studentov t-test. Koristi se za testiranje značajnosti razlika između dve aritmetičke sredine.

Uslovi za primenu t-testa:

1. Obe varijable koje se testiraju moraju biti numeričke.
2. Ukoliko je veličina uzorka manja od 30 jedinica, raspored treba biti normalan ili bar simetričan.

Za njegovo realizovanje potrebno je poznavati parametre statističkog skupa: veličinu uzorka (n), standardnu devijaciju (SD), i aritmetičku sredinu ($\bar{X}$).

Nije potrebno poznavanje varijanse osnovnog skupa, pa je ovaj tip testa praktičniji od z – testa, jer se testiranje hipoteze o aritmetičkoj sredini osnovnog skupa najčešće odvija u uslovima kada je varijansa osnovnog skupa nepoznata. U tim uslovima varijansu osnovnog skupa procenjujemo na osnovu varijanse uzorka, odnosno grešku ocene aritmetičke sredine osnovnog skupa izračunavamo na osnovu standarne devijacije uzorka po obrascu:

$$SG = \frac{SD_{uz}}{\sqrt{n-1}}$$

gde je n-1 stepen slobode. Pod uslovom da osnovni skup uma normalan raspored ili da je n>30, a varijansa osnovnog skupa nije poznata, testiranje hipoteze zasniva se na statistici Studentovog t-testa, koji se izračunava po obrascu:

$$t = \frac{\bar{X}_{uz} - \bar{X}_{os}}{\frac{SD_{uz}}{\sqrt{n-1}}}$$

gde je X osnovnog skupa hipotetična, unapred poznata vrednost.

Ako je realizovana t-vrednost manja od granične tablične vrednosti za odgovarajući broj stepena slobode i prag značajnosti, nulta hipoteza se prihvata kao tačna, a odbacuje alternativna hipoteza. Obrnuto, ako je realizovana t-vrednost jednaka ili veća od granične tablične vrednosti, za odgovarajući broj stepena slobode i prag značajnosti, nulta hipoteza se odbacuje kao netačna, a prihvata se alternativna hipoteza.

19. Man-Vitnijev test

Ovaj test se primenjuje za testiranje hipoteze o jednakosti neprekidnih raspodela za obeležja X i Y na osnovu dva slučajna uzorka ($X_1, X_2, \ldots, X_m$) i ($Y_1, Y_2, \ldots, Y_n$) pri čemu je n≥m. Efikasniji je od t – testa (0,95) kod raspodela koje su različite od normalne i slične efikasnosti kod normalne raspodele, pri čemu je:

$$H_0(F_x(x) = F_y(x))$$

$$H_1(F_x(x) \neq F_y(x)).$$

Pri testiranju se formira objedinjeni uzorak sortiran u neopadajućem poretku.

Jedan od načina da se izračuna vrednost statistike U je da se saberu svi rangovi elemenata X i svi rangovi elemenata Y. Tada se vrednost test – statistike računa na osnovu jedne od ove dve formule:

$$U = R_x - \frac{m(m+1)}{2}$$

$$U = -R_y + mn + \frac{n(n+1)}{2}$$

$$U_1 + U_2 = mn$$

Aproksimacija normalnom raspodelom je dobra već za m,n ≥ 8. Ako su obimi uzoraka manji od 8, koriste se posebne tablice.

20. Analiza varijanse (ANOVA)

Često je potrebno porediti i više od dve grupe različitih ispitanika (grupe različitih sportista, odeljenja u školi, klubova u nekom takmičenju i sl.). U takvim slučajevima koristi se statistička metoda poznata kao Analiza varijanse ili ANOVA (engl. *Analysis of Variance*). Izbor ispitanika u grupama treba da bude slučajan i nezavisan, a varijabiliteti rezultata u populacijama analiziranih grupa treba da budu statistički jednaki. Rezultati grupa ispitanika treba da budu normalno distribuirani, odnosno da ne odstupaju statistički značajno od normalne distribucije.

Osnovna logika analize varijanse sastoji se u tome da se testira odnos varijabiliteta rezultata između grupa i varijabiliteta unutar grupa ispitanika.

Ako se analizira položaj nekog rezultata (X) u masi svih rezultata, može se zaključiti da se on sastoji iz dve komponente:

1. **Varijabiliteta unutar grupe** – odstupanja u odnosu na aritmetičku sredinu svoje grupe.
2. **Varijabiliteta između grupa** - odstupanja aritmetičke sredine kojoj pripada rezultat od zajedničke aritmetičke sredine.

U analizi varijanse važna je tzv. suma kvadrata odstupanja rezultata od odgovarajuće aritmetičke sredine, odnosno varijansa.

Za sve ispitanike suma kvadrata je:

$$SS_T = \sum_{i=1}^{n} (X_i - AS_{tot})^2$$

Suma kvadrata unutar grupa je:

$$SS_{ug} = \sum_{i=1}^{N_g} (X_{ig} - AS_g)^2$$

Suma kvadrata između grupa je:

$$SS_{bg} = \sum N_g \left(AS_g - AS_{tot}\right)^2$$

Odnos svih suma kvadrata je:

$$SS_T = SS_{bg} + SS_{ug}$$

Testira se nulta hipoteza $H_0 : AS_g = AS_{tot}$, odnosno da je varijabilitet oko zajedničke aritmetičke sredine (MS_b) statistički jednak varijabilitetu oko aritmetičkih sredina grupa (MS_u). To testiranje vrši se **F**-odnosom:

$$F = \frac{MS_b}{MS_u}$$

Kada je nulta hipoteza odbačena, tada se može računati t-testom između kojih parova grupa postoji statistički značajna razlika. F-odnos ili F test, ima očekivanu vrednost 1, a veća vrednost od određene granične vrednosti ukazuje na postojanje statistički značajne razlike između analiziranih grupa ispitanika na posmatranoj varijabli.

U istraživačkoj praksi se često javlja potreba da se za neke grupe ispitanika testiraju razlike na osnovu dve, pa i više nezavisnih, faktor varijabli. Za takve analize se koristi posebna varijanta analize varijanse koja se naziva **dvofaktorska analiza varijanse** (eng. *Two-Way ANOVA*). Tada se testiraju tri nulte hipoteze: da razlike za prvi faktor nisu statistički značajne, da razlike za drugi faktor nisu statistički značajne i da interakcija ovih faktora nije statistički značajna.