Dr Đurđica Vasilj,
Poljoprivredni fakultet — Zagreb

ULOGA TRANSFORMACIJA U ANALIZI VARIJANCE

Najpoznatija i najčešće primjenjivana tehnika u statističkoj obradi eksperimentalnih podataka je sigurno analiza varijance.

Međutim, treba imati na umu, da su rezultati (a pogotovo testovi signifikantnosti) koji se pomoću nje dobiju, valjani samo onda ako podaci zadovoljavaju određene uvjete — postavke.

Stoga je od neobične važnosti poznavati te postavke, poznavati metode pomoću kojih se može testirati da li su postavke zadovoljene, znati posljedice do kojih dolazi ako su one narušene, odnosno znati što treba poduzeti da bi se one eventualno zadovoljile.

Treba naglasiti, da se prije provođenja bilo kakve analize varijance na podacima nekog stvarnog problema, mora najprije provjeriti da li su postavke zadovoljene. Ako one to nisu, treba poduzeti jedan od nekoliko mogućih koraka za ispravljanje tog odstupanja.

Na sreću, u vrlo mnogo slučajeva se jednostavno promjenom skale mjerenja, tj. transformiranjem osnovnih podataka u neku drugu skalu, može postići zadovoljavanje potrebnih postavki.

Međutim, kad se ni tranformacijama ne može podatke učiniti takvim da zadovoljavaju postavke, treba znati primijeniti i neke druge statističke tehnike. Jedna od njih je upotreba neparametrijskih metoda (Vasilj: Agr. glasnik br. 8—9/1972; tamo je naglašeno da se one mogu koristiti čak i u slučajevima kad su postavke za analizu varijance zadovoljene, a žele se samo dobiti što brži rezultati. U tom se slučaju gubi na točnosti i valjanosti u odnosu na analizu varijance koja je mogla biti provedena).

Najvažnije postavke na kojima bazira analiza varijance su: da je uzorak slučajan, da su efekti tretiranja kao i okolinski efekti aditivni, da su pogreške pokusa neovisne i raspoređene prema normalnoj distribuciji, te da su varijance homogene.

Najosnovnije je dakle osigurati stvaranje slučajnog uzorka. U protivnom dolazi do narušavanja postavke o neovisnosti pogrešaka, ili do heterogenosti varijanci, ili do raspodjele koja ne slijedi normalnu.

Neovisnost pogrešaka u uzorku postiže se slučajnim uzimanjem podataka (odnosno tzv. randomiziranim rasporedom tretiranja) kao i ponavljanjima tretiranja ili repeticijama. Treba naglasiti da je jako važno o svemu ovome voditi računa već kod postavljanja pokusa (slučajni raspored, repeticije), jer se kasnije ništa od ovog ne može popraviti ako je promaklo kod postavljanja pokusa. Ako ova postavka, tj. da su pogreške slučajne i neovisne nije zadovoljena, to može jako narušavati valjanost uobičajenog »F« testa.

Jednakost (homogenost) varijanci u grupi uzoraka je važan preduvjet za niz statističkih testova. Postoje različiti testovi za testiranje homogenosti varijanci (»F« test, Bartlettov test itd.), a kako se pri analizi pretpostavlja da je svaka varijanca tek procjena prave varijance pogreške, to njihovo provođenje i te kako ima smisla. Razlozi za heterogenost varijanci mogu biti različiti (neki speciesi su jače varijabilni u nekom svojstvu nego drugi i sl.), a vrlo često može razlog tome biti krivi izbor skale mjerenja. U nekim sklama mjerenja varijance su funkcije srednjih vrijednosti. Kod varijabli koje recimo slijede Poissonovu distribuciju (a to je distribucija rijetkih slučajeva), varijanca je jednaka srednjoj vrijednosti, pa će populacije s većim prosjekom imati i veću varijancu. Ovakvo odstupanje od homogenosti može biti lako popravljeno upotrebom nove skale mjerenja — transformiranjem podataka. Odstupanje od aditivnosti se također može izbjeći pomoću transformacija. Svaka opažena (izmjerena) vrijednost može naime biti raščlanjena u aditivne komponente, koje predstavljaju efekti tretiranja (u izvjesnom redu ili stupcu u dvosmjernoj analizi varijance) i slučajni efekti. Ako se radi o multiplikativnoj a ne aditivnoj situaciji (što se već iz podataka može zaključiti), transformiranjem podataka u pogodnu novu skalu, može se ipak postići aditivnost.

U praksi međutim često sve te postavke ne stoje, no važno je znati posljedice do kojih dolazi uslijed odstupanja od njih. Naime, nezadovoljavanje ovih osnovnih postavki za analizu varijance djeluje na granice opravdanosti i osjetljivost »F« i »t« testa. Zato se može dogoditi da mislimo da testiramo na granici od 5 %, a zapravo testiramo na recimo 8 % i sl. Na taj način se u analizi varijance mogu činiti pogreške poznate kao pogreška tipa I i pogreške tipa II. Ako se na temelju uzorka testiranjem utvrdi opravdana razlika a ona to u stvari nije, učini se pogreška tipa I i obratno pogreška tipa II. Drugim riječima odbacivanje istinite nul-hipoteze dovodi do pogreške tipa I, a prihvaćanje lažne nul-hipoteze do pogreške tipa II. Često se tako dobije premalen ili prevelik broj opravdanih ili neopravdanih rezultata (u odnosu na stvarno stanje). Naravno da uz sve ovo dolazi isto tako i do razmjernog gubitka točnosti u procjenama efekata tretiranja (a koji se mogu točnije procijeniti ako se zna točan model kojeg treba slijediti). Do najozbiljnijih poremetnji dolazi ako varijanca pogreške eksperimenta nije konstantna za sva tretiranja. U tom slučaju odgovarajuća varijanca pogreške za kompariranje jednog para tretiranja može biti nekoliko puta veća nego ona za drugi par i zato bi uzimanje iste pogreške za testiranje oba para dovelo do potpuno pogrešnog »t«-testa. U takvom slučaju treba pogrešku podijeliti u komponente koje su homogene.

Do istog problema dolazi ako pogreške eksperimenta slijede neku iskrivljenu (asimetričnu i sl.) distribuciju, jer tada varijanca pogreške tretiranja postaje funkcija prosječne vrijednosti (srednje vrijednosti) tretiranja.

Ako je poznata priroda funkcionalne veze može se naći transformacija koja će podatke svesti na skalu u kojoj je varijanca pogreške konstantna (ili bar konstantnija). Svakako, transformacija se napravi prije početka analize.

Dakle, poznavanje osnovnih postavki za analizu varijance, posljedica koje nastaju ako one nisu zadovoljne i načina na koje se te postavke nekad mogu zadovoljiti, osigurava provođenje točnijih testova i zato donošenje

valjanih zaključaka na temelju podataka provedenih eksperimenata. Kako je već spomenuto, jedan od načina na koji se često može postići zadovoljavanje postavki o neovisnosti pogrešaka, aditivnosti i normalnoj raspodjeli, je transformiranje originalnih podataka u neku drugu skalu. To je uloga transformacija.

Treba naglasiti da je za sve transformacije značajno da relativni odnosi (i poredak) originalnih podataka kao i njihovih prosječnih vrijednosti ostaju nepromijenjeni. Stoga ako su dvije srednje vrijednosti dobivene iz originalnih podataka iste, njihove transformirane vrijednosti će također biti iste.

Teško je dati neko opće pravilo za izbor vrste transformacije, jer primijenjena transformacija možda neće odstraniti sve nepoželjnosti u podacima. Međutim, izvjesne se transformacije preporučuju kao najpogodnije za svojstva izražena u vidu postaka jedne vrste, neke pak za postotke druge vrste, neke za svojstva koja brojčano izražavaju učestalost individua na izvjesnoj površini (tla, petrijevke i sl.), druge pak za podatke koji slijede Poissonovu distribuciju (mali cijeli brojevi) itd. Zato je potrebno najprije dobro pogledati podatke (najbolji je način prikazati ih u vidu grafa i promatrati njihove distribucije), pa tek onda odlučiti da li upotrebiti transformaciju i koju.

U ovom će prikazu biti opisane neke transformacije koje se najčešće primjenjuju, a koje doprinose stabilizaciji varijance i približavanju normalnoj distribuciji.

U literaturi su mnogi autori opisivali upotrebu transformacije za različite tipove podataka, a najčešće se nalaze radovi Bartletta, Bealla, Elissa, Cochrana, Williamsa, Powersa i dr. U nas se je na tome radilo vrlo malo, pa je svrha ove rasprave ukazati na važnost i opravdanost primjene bar nekih najvažnijih transformacija .

Na najjednostavnijem općem primjeru može se dobro uočiti što zapravo učini transformacija. Evo kako: svaka varijanta (u najjednostavnijem slučaju — jednosmjernoj klasifikaciji) sastoji se od $Y_{ij} = \mu + \alpha_i + E_{ij}$ gdje je μ = srednja vrijednost, α_i su efekti tretiranja, a Eij su pogreške. U ovakvom modelu očito su komponente aditivne, a pogreške normalno raspoređene. Međutim, u slučaju da su komponente multiplikativne, tada je

$$Y_{ij} = \mu \alpha_i E_{ij}$$

Tu je jasno narušena postavka o aditivnosti efekata i homogenosti varijanci. Opća srednja vrijednosa μ je konstantna u svakom slučaju, no efekti tretiranja i pogreške se razlikuju od grupe do grupe. Ovo možemo ispraviti jednostavno transformirajući dani model u logaritme, pa dobijemo

$$\log Y_{ij} = \log \mu + \log \alpha_i + \log E_{ij}$$

što predstavlja aditivni model. Na tako transformiranim podacima će se moći provoditi valjana analiza varijance. Najčešće time što podatke učinimo takvima da su varijance jednake, ujedno zadovoljavamo postavku o aditivnosti i normalnoj raspodjeli.

Jednostavan način na koji se može utvrditi da li će izvjesna transformacija dovesti do zadovoljenja uvjeta potrebnih za analizu varijance je da se naprave grafovi. Promjenom linearne skale u neku drugu skalu mjerenja, lako se može uočiti da li se krivulja osnovnih podataka (koja je prethodno pokazivala iskrivljenost — asimentričnost u jednom ili drugom pravcu) ispravila i da li zadovoljava uvjet o normalnoj distribuciji. Razumljivo je da se izabere ona transformacija koja distribuciju osnovnih podataka učini najsličnijom normalnoj. Nakon izbora transformacije, osnovni podaci se transformiraju dakle u novu skalu i analiza varijance kao i svi testovi izvrše se na osnovu transformiranih podataka.

1. $\sqrt{x}$ i $\sqrt{x + \frac{1}{2}}$ transformacija

Kad se podaci sastoje od cijelih brojeva dobivenih prebrojavanjem (kao broj insekata na listu, broj bakterijskih kolonija u petrijevki, broj biljaka na izvjesnoj površini), homogeni uvjeti često dovode do distribucije koja je slična Poissonovoj (a ne normalnoj), a sadrži podatke u vidu malih cijelih brojeva. Kako je kod takve distribucije varijanca jednaka srednjoj vrijednosti, a to znači da one nisu neovisne nego variraju ovisno jedna o drugoj, to nije zadovoljen jedan od osnovnih uvjeta za analizu varijance. Transformiranjem podataka u vrijednosti njihovog drugog korijena $(\sqrt{x})$ dobije se normalna distribucija i stabilizira se varijanca.

Često su podaci izraženi vrlo malim brojevima (manjim od 10) pa čak i nulama. U takvom se slučaju preporučuje upotrebiti $\sqrt{x + \frac{1}{2}}$ transformaciju. Prije transformiranja se naime svakom podatku doda 0,5 pa se onda izvrši transformacija. Ako se kod prikazivanja rezultata želi upotrebiti originalnu skalu mjerenja, što je najčešće poželjno radi jasnoće, dobivene transformirane vrijednosti se jednostavno kvadriraju. Ova se dakle transformacija može preporučiti kod svojstava izraženih podacima u vidu malih cijelih brojeva dobivenih prebrojavanjem u nekakvoj homogenoj sredini koja je uvjetovala odstupanje od normalne distribucije. Pogodna je i za podatke u vidu postotaka (dobivenih brojenjem i na bazi istog nazivnika) koji se kreću od 0 do 20 % i 80 do 100 % ali ne oboje. Kod toga se vrijednosti između 80 % i 100 % najprije odbiju od 100 a onda transformiraju.

2. Logaritamska transformacija

Moglo bi se reći da se ova transformacija najčešće upotrebljava u svim onim slučajevima gdje je srednja vrijednost u pozitivnoj korelaciji s varijancom (znači većoj srednjoj vrijednosti odgovara veća varijanca). Logaritamska transformacija normalizira distribuciju i stabilizira varijancu. Jako često se ova transformacija primjenjuje u regresionoj analizi, gdje se najčešće transformiraju podaci ovisne varijable. Primjenjuje se na podatke koji se sastoje od brojeva jako širokog raspona. Kada podaci sadržavaju vrijednost nula, treba svakoj vrijednosti prije transformiranja dodati jedinicu pa onda transformirati, odnosno log $(x + 1)$, jer je $\log 0 = -\infty$. Kad se vrijednosti kreću između 0 i 1 poželjno je prije transformiranja svaku vrijednost po-

množiti s 10, 100 ili bilo kojom potencijom od 10 da se izbjegnu negativni predznaci logaritama. Za ovu transformaciju se najčešće koriste dekadski logaritmi (iako se mogu upotrebiti logaritmi na bilo kojoj bazi). Pri tome je u obradi podataka dovoljno koristiti 2—3 mjesta mantise logaritma, pogotovo kad se radi sa stolnim kalkulatorima.

Logaritamska transformacija se najčešće koristi za podatke o svojstvima kao što je rast organizama izražen u vidu dimenzije porasta (porast duljine i dr.).

3. Kutna, inverzna sinus transformacija ili arc sin transformacija

Podaci izraženi u obliku proporcija odnosno postotaka, često predstavljaju izvore narušavanja osnovnih postavki za analizu varijance. U literaturi se za takve podatke preporučuje arc sin transformacija (ili tzv. kutna transformacija ili inverzna sinus transformacija). Prema tome bi podaci izraženi u postocima (pogotovo ako su jako širokog raspona) trebali u pravilu biti podvrgnuti transformaciji. Kod toga se svaka vrijednost (x) transformira u kut čiji je sisnus x. Termin arc sin je dakle sinonim za inverznu vrijednost sin ili $\sin^{-1}$, odnosno arc sin $\sqrt{p} = \theta$ gdje je p proporcija (ili postotak).

Općenito se uzima da se transformirati trebaju samo podaci koji se kreću unutar raspona 0—30 % i 70—100 %. Ako su podaci izraženi postocima od 30 do 70 % obično ih nije potrebno transformirati (jer su u većini takvih slučajeva zadovoljene postavke za analizu varijance). Ovdje treba napomenuti, da je vrlo važno o kakvoj se vrsti postotaka radi. Razlika je naime u vrsti postotka koji izražava recimo koncentraciju i onog koji izražava % klijavosti. U prvi tip spadala bi i svojstva kao: čistoća sjemena (izražena kao težina čistog sjemena/ukupna težina sjemena), pa sadržaj proteina (težina proteina/ukupna težina) ili sadržaj šećera u % itd. I prirod može recimo biti izražen kao postotak od kontrole (ili standarda) umjesto stvarnog priroda u kg, q i sl.

Drugi tip postotaka predstavljaju svojstva kao % klijavosti ili % zaraze i sl. To su podaci koji su bazirani na nekom određenom broju slučajeva (n). Ove podatke, za razliku od prvih treba gotovo u pravilu transformirati prije provođenja analize varijance.

Arc sin vrijednosti su izražene u stupnjevima ili u radijanima, a njihova je varijanca konstantna i iznosi $\frac{821}{n}$ ako su u stupnjevima ili $\frac{0{,}25}{n}$ ako su u radijanima. Pri tom je n nazivnik, zajednički za sve proporcije.

U praksi se međutim često događa da n nije sasvim isti za sve proporcije. Ako je on vrlo sličan, transformacija je ipak dozvoljena i korisna.

U slučajevima gdje su podaci predstavljeni vrijednostima koje sadrže i nulu (0 %) ili 100 %, Bartlett (1) preporučuje da se prije transformiranja nule zamijene sa $\frac{1}{4n}$, a stotice sa $100 - \frac{1}{4n}$.

Kako je već naprijed spomenuto, podaci u vidu proporcija ili postotaka mogu nekad biti transformirani i u logoritamske vrijednosti u svrhu postizanja uvjeta za analizu varijance. Međutim, najčešće je arc sin transformacija najdjelotvornija.

4. Recipročna transformacija

U biološkim istraživanjima često se ispituju korelativni odnosi između svojstava kao: broj udaraca krilima u sekundi (kod insekata) ili fertilitet u vidu broja snešenih jaja po ženki i sl. Predstavljeni grafom takvi podaci obično daju krivulje koje se matematički mogu opisati kao bxy = 1 ili
(a + bx) y = 1.

Iz toga se može izvesti $\frac{1}{y} = bx$ i $\frac{1}{y} = a + bx$.

Transformacijom ovisne varijable u njezinu recipročnu vrijednost često možemo tako umjesto krivulje dobiti pravac.

5. Probit transformacija

U mnogim istraživanjima u biologiji tretiraju se problemi o djelovanju insekticida i pesticida uopće, ispituje se djelovanje droga, seruma i sl. U svim tim problemima je djelovanje izvjesnog tretiranja izraženo u vidu reakcije živih organizama na primijenjeni tretman. Obično se tretiranja sastoje u primjeni različitih doza na izvjesne grupe individua i praćenju njihovih reakcija. Tu se zapravo radi o dvije varijable čije distribucije se normaliziraju primjenom logaritamske i probit transformacije.

Da se lakše uoči bit ove transformacije evo konkretnog primjera: ako se određuje toksičnost nekog preparata za izvjesnu grupu insekata, obično se različitim koncentracijama tretira određen broj individua i brojenjem utvrdi smrtnost. Rezultati mogu biti izraženi kao proporcije m/n (m = broj uginulih, n = ukupni broj insekata) ili kao postoci 100 m/n. Upravo takvi podaci su podložni tzv. probit analizi, koja uključuje i probit transformaciju. Kod djelovanja nekog insekticida postoji teoretski jedna granica koncentracije iznad koje insekti ugibaju, a ispod koje preživljavaju. Taj se nivo naziva tolerancijom (λ). Općenito se može reći da se za većinu preparata distribucija ne podudara s normalnom, ali se može jako približiti normalnoj ako se vrijednosti transformiraju u logaritme ($x = \log \lambda$). Na taj se način operira s logaritamskim vrijednostima primjenjenih koncentracija (neovisne varijable) raspoređenih prema normalnoj distribuciji.

Prema Finney-u (10) λ se naziva dozom u smislu stvarne koncentracije (kao mg/l i sl.), a logaritamska vrijednost doze zove se »doziranje« (engl. dosage ili dose metameter).

Treba naglasiti da log transformacija ne normalizira baš uvijek distribuciju tolerancija. Nekad se u tu svrhu mogu upotrebiti i druge transformacije, ali dokazano je da je u najvećem broju slučajeva djelotvorna baš log transformacija.

Ispitujući tako djelovanje različitih doza (koncentracija) nekog insekticidacida, podaci o broju uginulih individua (ovisno o primjenjenoj dozi) mogu se prikazati u vidu grafa na čijoj apscisi se označe logaritamske vrijednosti koncentracije ili $x = \log \lambda$ (koje su tada normalno distribuirane), a na ordinati postotak uginulih insekata.

Graf u pravilu predstavlja krivulju-sigmoidu. Ovakvi podaci se mogu analizirati pomoću regresione analize, ako se vrijednosti o % uginulih individua transformiraju u novu skalu. Na taj se način umjesto sigmoidne krivulje dobija pravac, s kojim je onda jednostavno raditi.

Transformiranje postotaka, na način da transformirane vrijednosti leže u odnosu na koncentracije na pravcu, zapravo je smisao probiti transformacije. Postoci se naime transformiraju u tzv. N. E. D. vrijednosti (normal equivalent deviation). To su zapravo odstupanja svake log vrijednosti koncentracije od prosjeka, izražene u dijelovima standardne devijacije

$$\text{N.E.D.} = \frac{x - \bar{x}}{\delta}$$

Da bi se izbjegle negativne vrijednosti, Bliss (2) je predložio da se svakoj N.E.D. vrijednosti doda 5 i tako dobivene vrijednosti nazvao probitima (probit = N.E.D. + 5).

Primjenom log i probit transformacija normaliziraju se distribucije, a time se uglavnom zadovoljavaju i ostali uvjeti za analizu varijance, odnosno računanje regresije ako se radi o odnosu dviju varijabli.

Kako je probit analiza jedno posebno i veliko područje u biometričkoj analizi, to se onog čitaoca koji je želi još detaljnije upoznati može uputiti na autore: Gouldena (11), Finneya (10), Busvinea (5), koji je opisuju u detalje.

Ovim prikazom obuhvaćeni su samo najvažniji tipovi transformacija koji dolaze najčešće u obzir u statističkoj obradi eksperimentalnih podataka.

Gotovo u svakoj statističkoj knjizi naći će se potrebne tablice (drugih korijena, logaritama, recipročnih vrijednosti, arc sin $\sqrt{\%}$ i probita), pa ih zbog opsežnosti nema smisla ovdje navoditi.

Važno je još jednom naglasiti da je kod analiziranja podataka koji ne slijede normalnu distribuciju, upotreba transformacija prije analize jedan od najtočnijih puteva koji vode do valjanih testova i zaključaka.

Svakako, od presudnog značenja je pravilan izbor transformacije. Zato treba najveću pažnju obratiti upravo detaljnom proučavanju osnovnih podataka, a tek onda izabrati i odlučiti koja transformacija će biti najdjelotvornija.

THE RULE OF TRANSFORMATIONS IN THE ANALYSIS OF VARIANCE

SUMMARY

The assuptions for the analysis of variance and use of transformations were discussed. The main assumptions were listed (random variables, additivity, homogenity of variances, normality and lack of correlations), and some consequences when the assumptions for the analysis of variance are not satisfied were described.

The use of various transformations as related to variance stabilization and normalization was discussed. Those were:

1. Square root transformation
2. Log transformation
3. Arc sin transformation
4. Reciprocal transformation
5. Probit transformation

It was emphasized that no one transformation works perfectly. As a consequence it is very important to decide wheather one should use a certain type of transformation.

LITERATURA

Bartlett, M. S. (1947): The use of transformations. Biometrics 3, 39—52.

Bliss, C. I. (1934): The method of probits. Science 79, 409—410.

Bliss, C. I. (1935): The calculation of the dosage-mortality curve. Ann. Appl. Biology 22, 134—167.

Bliss, C. I. (1954): An outline of biometry. Yale Co-op Corp., New Haven, Connecticut.

Busvine, J. R. (1957): A critical review of the techniques for testing insecticides.

Cochran, W. G. (1943): Analysis of variance for percentrages based on unequal numbers. Journ. Am. Stat. Assoc. 38, 287—301.

Cochran, W. G. (1947): Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics 3, 22—38.

Clark, A. G. and W. H. (1939): The analysis of variance with special reference to data expressed as percentages. Journ. Am. Soc. Agr. 31, 55—67.

Dolby, J. A (1963): A quick method for choosing a transformations. Technometrics 5, 317—325.

Finney, D. J. (1952): Probit analysis. Rev. ed. Cambridge Univ. Press, London.

Goulden, C. H. (1952): Methods of statistical analysis. Wiley, New York.

Powers, L. R. (1950): Determining scales and the use of transformations in studies of weight per locule of tomato fruit. — Biometrics 6, 145—163.

Sokal, R. R. and F. J. Rholf (1969): Biometry. The principles and practice of statistics in biological research. W. H. Freeman and Co., San Francisco.