Pregledni rad

UDK 519.852: 339.137.025

DOI 10.7251/SVR1918279M

SIGNIFIKANTNOST KOEFICIJENTA LINEARNE KORELACIJE

Doc. dr Branka Marković[1]

Prof. dr Marinko Markić[2]

Nezavisni univerzitet Banja Luka

Apstrakt: U prirodi i društvu postoji određena pravilnost i zakonitost između pojava. Odnos koji postoji može se definisati kao funkcionalan i statistički. Kod funkcionalnog odnosa vrijedi pravilo, za koliko se mijenja jedna varijabla, isto toliko se mijenja i druga varijabla.Kod funkcionalne veze se tačno utvrđuje vrijednost jedne za datu vrijednost druge pojave. Statistička veza u odnosu na funkcionalnu, pokazuje koja je jako čvrsta veza , ova je slaba veza , kod koje promjene nezavisno promjenljive na izazivaju promjene zavisno promjenljive varijabле. Statistička veza je specifična veza i javlja se kod društvenih pojava. U statističkoj analizi utvrđuje se odnos između dvije ili više pojava. Analiza predstavlja prikaz i utvrđivanje brojčanih pokazatelja i izraza kojim se u datom momentu omogućava donošenje odluke o samoj pojavi.

Osnovni skup, u većini slučajeva, nije homogen u odnosu na posmatrano obilježje. On je po pravilu heterogen i sastavljen je od više heterogenih podskupova. Ova podjela se zove stratifikacija, a podskupovi koji se dobiju na ovaj način zovu se blog ili stratum. Podskup iz nepodjeljenog skupa zove se jednostavan podskup. Podaci iz tog podskupa predstavljaju uzorak. Uzorak je uvijek manji od osnovnog skupa i čini osnovu za zaključivanje o osnovnom skupu. Po svojim karakteristikama uzorak treba da odražava karakteristike osnovnog skupa. Pošto uzorak sadrži samo dio elemenata osnovnog skupa, a ne sve podatke, analitički pokazatelji će da sadrže grešku koja je posljedica podskupova podataka. Upravo zbog ovih grešaka i karakteristika uzorka, potrebno je izvršiti testiranje kojim se provjerava postavljena hipoteza nepoznatog osnovnog skupa na osnovu uzorka.

Ključne riječi: *uzorak, osnovni skup, korelacija, koeficijent*

UVOD

Ispitivanje zavisnosti između dvije pojave vrši se korištenjem χ^2testa, te pomoću Pirsonovog koeficijenta kontigencije, kojim se određuje jačina veze između dvije pojave. Obje analize imaju određene nedostatke: vrijednost jedne varijable se ne može odrediti na osnovu vrijednosti druge varijable;ne može se odrediti oblik veze koji postoji između pojava;ne može se odrediti zavisnost između dvije pojave, a ne između više pojava.

[1] NUBL Banja Luka

[2] NUBL Banja Luka

Kao dio statističke analize , koju smatraju kao najvažniju, kojom se analizira međuzavisnost dvije ili više varijabli predstavlja regresija i korelacija.Istražuje se postojanje veze između varijabli i njihova jačina. Stepen statističke povezanosti određuje se pomoću metode korelacione analize, dok regresijska metoda ima zadatak da utvrdi oblik veze, zavisnost između varijabli. Regresijski model je algebarski model koji korištenjem matematičke jednačine opisuje kvantitativnu zavisnost između posmatranih varijabli.Iako su ova dva modela u bliskoj povezanosti, ipak postoji određena razlika između njih. Osnovna razlika je u tome, što je kod regresijskog modela potrebno unaprijed odrediti koja pojava je zavisno, a koja nezavisno promjenljiva, dok kod korelacionog modela nije bitno, pri analizi dvije pojave koja je zavisna a koja nezavisna varijabla, ali kod korelacione analize tri i više pojava potrebno je odrediti koja je varijabla zavisno promjenljiva. Opšti model regresijske analize glasi:

$$Y=f(x)+u$$

Pri čemu je *y* zavisna varijabla(regresand varijabla), *x* nezavisna varijabla(regresorska varijabla), dok varijabla *u* predstavlja nepoznato odstupanje od funkcionalne veze. Pojava varijable *u* ukazuje na prisustvo statističkog odnosa između pojava. Ukoliko se regresijski model sastoji samo od jedne zavisne i jedne nezavisne govori se o jednostavnoj regresiji, a može da ima dvije ili više varijabli u tom slučaju kažemo da se radi o višestrukoj regresiji.

Regresijska i korelaciona analiza sprevode se na osnovu stvarnih vrijednosti pojave. Svi statistički podaci u modelu regresijske analize su numerički izražene vrijednosti varijabli u modelu. Pored stvarnih vrijednosti varijabli u regresijskoj analizi javljaju se i indikator varijable, čija je vrijednost jednaka 0 i 1. Indikator određenih varijabla se javlja kod nominalnih obilježja.

Kako je već objašnjeno u prethodnom dijelu, osnovni zadatak statističke analize je utvrđivanje odnosa između pojava i numeričko izražavanje stepena njihove povezanosti i predstavljanja pomoću algebarskog izraza tj. regresijskim modelom. Numerička analiza stepena povezanosti, jačina veze između pojava izražava se pomoću jednostavne, parcijalne ili višestruke korelacije, koeficijenta korelacije ranga i koeficijenta asocijacije. Kod istraživanja pojava utvrđuje se koeficijent korelacije a zatim se nastavlja sa regresijskom analizom. Zavisno od primarnog cilja analize, ako je cilj istraživanja predikatni oblik, istraživanje polazi od regresijskog modela.

1.0 Korelaciona analiza

Pošto se polazi od statističke prirode odnosa, gdje su veze između pojava mnogo slabije, nego kod funkcionalnih, zadatak je mjerenje stepena kvantitativnog slaganja (kovarijacije varijabli). Upravo, mjerenje stepena jačine statističke veze između dvije ili više varijabli vrši se metodom korelacione analize. Pokazatelj stepena jačine statističke veze između dvije pojave utvrđuje se koeficijentom jednostavne linearne korelacije.

1.1.Koeficijent jednostavne linearne korelacije

Koeficijentom jednostavne linearne korelacije se mjeri stepen i smjer povezanosti dvije pojave u linearnom odnosu. Kod linearnog statističkog odnosa rasipanje, koje se prikazuje pomoću dijagrama, je duž zamišljenog pravca. Mjera jačine proste linearne korelacije je relativna mjera.

1.1.2. Pearsonov koeficijent korelacije

Temelj za izračunavanje Pearsonovog koeficijenta korelacije je kovarijansa. Kovarijansa predstavlja prvi mješoviti moment oko sredine i predstavljen je izrazom:

$$\mu_{11} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n} \qquad (1.1)$$

Pošto se u brojniku izraza 1.1 nalazi proizvod odstupanja vrijednosti varijable *x* i *y* od njihovih aritmetičkih sredina. Vrijednost kovarijanse zavisi od tog proizvoda, ukoliko je jedan od faktora proizvoda jednak 0, za svaku vrijednost *i*, brojnik je jednak nuli, a time i kovarijansa je jednaka 0. Polazna osnova za računanje varijanse, je odstupanje vrijednosti varijable od njene aritmetičke sredine, pa time i vrijednost kovarijanse jednaka je 0, ukoliko je jedna varijansa jednaka 0. U praksi se ovakav slučaj vrlo rijetko javlja, pa je i vrijednost kovarijanse različita od 0. Predznak i vrijednost kovarijanse zavisi od predznaka brojnika.

Direktno mjerenje jačine linearne veze pomoću kovarijanse nije moguće, pošto kovarijansa zavisi od mjernih jedinica i veličine varijabli *x* i *y*. Zbog toga je potrebno eliminisati uticaj mjerne jedinice na varijable *x* i *y*, to je moguće ukoliko se vrijednosti odstupanja od njihove sredine izraze u jedinicama standardne devijacije. Pošto je aritmetička sredina standardiziranih vrijednosti jednaka 0, njihova kovarijansa se može se prikazati izrazom:

$$r=\frac{\sum_{i=1}^{n} z(x_i)z(y_i)}{n}$$

(1.2)

kada se izraz 1.2 uredi, dobije se Pearsonov koeficijent linearne korelacije koji glasi:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}} \qquad (1.3)$$

Pearsonov koeficijent predstavlja kovarijansu standardiziranih varijabli *x* i *y*.

Ukoliko je primarni cilj analize istraživanje predikatnog oblika, polazi se od regresijskog modela. Za određivanje koeficijenta korelacije koriste se rezultati regresijske analize, u takvom slučaju koeficijent linearne korelacije može se dobiti korištenjem koeficijenta determinacije. U ovom slučaju koeficijent linearne korelacije jednak je korijenu iz koeficijenta determinacije.

$$r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1.4)$$

Predznak koeficijenta uslovljen je predznakom regresijskog koeficijenta *b*. Interval u kome se kreće koeficijent je od -1 do +1, s tim da je veže sa veličinom koeficijenta determinacije. Negativan predznak pokazuje smjer veze, a ne njegovu jačinu. Ukoliko se vrijednost koeficijenta kreće oko 1, linearna veza je jaka i obrnuto.

1.2. Signifikantnost koeficijenta linearne korelacije

Regresijska i korelaciona analiza podudara se sa analizom kod ocjenjivanja i testiranja hipoteza:potrebno je iz osnovnog skupa uzeti reprezentativan uzorak , na osnovu kojeg se donosi zaključak o parametrima osnovnog skupa.Na osnovu vrijednosti koeficijenta korelacije u uzorku daje se zaključak o koeficijentu korelacije osnovnog skupa.

Ukoliko se iz normalno raspoređenog osnovnog skupa izabere uzorak od *n* elemenata, kod svakog uzorka mjere se dva obilježja x_i i y_i i pri tome se izračuna koeficijent linearne korelacije uzorka. Koeficijent uzorka *r* je različit od nule, pri čemu su varijable uzorka linearno korelirane. Postavlja se hipoteza da su varijable x_i i y_i osnovnog skupa međusobno nezavisne, koeficijent linearne korelacije osnovnog skupa je jednak 0. Pri čemu se mora utvrditi da li uzorak obara hipotezu da u osnovnom skupu varijable nisu korelirane.U analizi je dokazano da statistika testa slijedi Studentov *t* raspored sa *n-2* stepena slobode.

$$t_0 = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \quad (1.5)$$

Potrebno je odrediti P vrijednost, koja predstavlja vjerovatnoću, pod predpostavkom da je nul hipoteza tačna, a da statsistika testa uzme ekstremnu vrijednost, koja je ekstremnija od one realizovane.

$$\mathrm{P}(|t| > t_0) = \bar{P}_{n_0}(t_0) = 2\{1 - S_{n_0}(t_0)\} \quad (1.6)$$

Pri čemu je $S_{n_0}(t_0)$ funkcija rasporeda Studentovog *t* rasporeda.

- Ukoliko je $\bar{P}_{n_0}(t_0) < 0{,}05$ odnosno $\bar{P}_{n_0}(t_0) < 0{,}01$, hipoteza se odbacije o nepovezanosti x_i i y_i osnovnog skupa, i dolazi se do zaključka da se koeficijent linearne korelacije *r* visoko signifikantno razlikuje od 0.
- Ako je $\bar{P}_{n_0}(t_0) > 0{,}05$ uzorak neće oboriti hipotezu nepovezanosti x_i i y_i osnovnog skupa i pri tome se koeficijent linearne korelacije *r* signifikantno ne razlikuje od nule nego slučajno.
- Ukoliko je n_0=*n-2*> 30, Studentov raspored se mijenja sa normalnim rasporedom.

ZAKLJUČAK

U statističkoj analizi korištenje metode korelacije ima veliki značaj pri testiranju hipoteza u naučnim radovima. Kod naučnih radova može se

primjetiti postojanje povezanosti dvije varijable i pokušava se utvrditi uzročno-posljedična veza koja se ne dokazuje korelacijom.

Kod upotrebe koeficijenta korelacije potrebno je znati na koji način se upotrebljava da bi se donijeli ispravni zaključci. Kod upotrebe korelacije, naročito u eksperimentalne svrhe, potrebno je upotrijebiti odgovarajuću matematičku funkciju koja je najbliža problemu koji se istražuje. Poslije čega se testiranjem provjerava zadana korelacija a pri tome se sama korelacija koristi za provjeru rezultata testiranja, nakon čega se utvrđuje korelacija između testiranja i dobijenih rezultata.

Nivo signifikantnosti se često pogrešno tumači. Nivo statističke signifikantnosti ne pokazuje jačinu veze između dvije varijable, nego pokazuje sa koliko povjerenja treba posmatrati dobijene rezultate. Na signifikantnost iznosa koeficijenta linearne korelacije utiče i veličina uzorka. Iz malih uzoraka ($n< 30$) izračunavaju se umjerene korelacije koje statistički nisu signifikantne na nivou pouzdanosti $p< 0,05$, ali vrlo male korelacije koje se računaju iz velikih uzoraka ($n> 30$) su statsistički signifikantne.

SIGNIFICANCE OF THE LINEAR CORRECTION COEFFICIENT

Branka Marković PhD, professor Marinko Markić PhD

Abstract: In nature and society there is a certain regularity and legality between phenomena. The existing relationship can be defined as functional and statistical. For a functional relationship the rule is valid, for how much a variable is changed, the same variable is changed by another variable. For a functional connection, the value of one for the second value of the second occurrence is exactly determined. Statistical relationship with functional, which is a very strong link, this is a weak link, in which the changes are independently altered to cause changes of the dependent variables. The statistical link is a specific connection and occurs in social phenomena. The statistical analysis determines the relationship between two or more occurrences. The analysis represents the representation and determination of numerical indicators and expressions that allow at the given moment to make a decision about the occurrence.

The basic set is in most cases not homogeneous in relation to the observed characteristic. He is usually heterogeneous and is composed of several heterogeneous subsets. This division is called stratification, and subclasses obtained in this way are called a blog or stratum. The subgroup of an unprivileged set is called a simple subset. Data from this subset represent a sample. The sample is always smaller than the basic set and makes the basis for the conclusion of the basic set. By its characteristics the sample should reflect the characteristics of the basic set. Since the sample contains only part of the elements of the basic set, not all the data, the analytical indicators will contain an error resulting from the data subsystem.Precisely because of these defects and sample characteristics, a test should be carried out to check the hypothesis of an unknown base sample based on the sample

Key words: *sample, basic set, correlation, coefficient*

LITERATURA

1. Wheleer A. J. & Ganji R. A. (2010). Introduction to Engineering Experimentation

2. http://www.crashkurs-statistik.de/spearman-korrelation-rangkorrelation/ (dostupno 8.01.2019)
3. http://matheguru.com/stochastik/272-korrelation.html (dostupno 9.01.2017)
4. https://ldap.zvu.hr/~oliverap/MetodeIstrazivanjaFT/11_Korelacija.pdf (dostupno 5.03.2019)
5. http://www.snz.unizg.hr/test/test4/datoteke/200605200146410.Korelacija.pdf (dostupno 6.04.2019)