UDK: 519/626
Originalni naučni rad

O REDUKCIJI PODATAKA U STATISTICI

Vesna JEVREMOVIĆ, Jovan MALIŠIĆ

REZIME

U radu se proučava redukcija podataka u statistici tako što se razmatraju mogućnosti rekonstrukcije vrednosti uzoračkih statistika, kao i testiranja statističkih hipoteza ako se na podatke iz prošlosti dodaju novi podaci. Pritom se, umesto svih podataka iz prošlosti, raspolaže samo izračunatim vrednostima nekih statistika.

Ključne reči: Uzorački momenti, koeficijent varijacije, koeficijent asimetrije, koeficijent spljoštenosti, medijana, test tačaka zaokreta, test Kolmogorova-Smirnova, χ^2-test.

1. UVOD

Pojave slučajnog karaktera proučavamo na osnovu nekih podataka (uzorka). Ova rečenica izgleda sama po sebi jasna, a opet, sadrži mnoga pitanja na koja treba odgovoriti, da bi se takav iskaz mogao prihvatiti kao tačan. Najpre, treba znati šta to podrazumevamo pod «pojave slučajnog karaktera»; onda, koje su to metode kojima ih «proučavamo» i najzad, koji su to podaci i kako su dobijeni.

Filozofsko pitanje o pojmu slučajnosti nije tema ovog teksta, kao ni metode za dobijanje reprezentativnih uzoraka, nego razmatranje ideja o čuvanju i upotrebi (statističkih) podataka.

Ideja 1: Ako čuvamo sve podatke, onda znamo sve o posmatranoj pojavi. Naravno, ne baš sve, jer je to što imamo samo jedan uzorak. Možemo takođe postaviti pitanje pod kojim uslovima su podaci dobijeni, zato što bi to moglo da utiče na vrednosti koje su zabeležene. Ako bismo, dakle, imali i podatke o nekim pojavama koje su usko vezane sa posmatranom, onda bi, verovatno, slika koju imamo o proučavanoj pojavi bila kompletnija.

Ideja 2: Ako čuvamo samo ključne podatke, onda opet znamo sve. Pri tom smatramo da su ključni podaci takvi da se samo na osnovu njih može da odredi raspodela obeležja na populaciji. Tada možemo, metodom Monte-Karlo da modeliramo vrednosti posmatrane pojave tj. da generišemo uzorke. Jasno je da se nikad u stvarno slučajnoj pojavi ne može ponoviti istovetan beskonačni niz opservacija. S druge strane, pošto radimo sa konačnim nizovima podataka, šanse da dobijemo istovetan niz postoje, i mogu se izračunati. Realno gledano mi ni tada NE možemo dobiti isto, ako su rezultati merenja realni brojevi, ali pošto rezultate uvek zaokružujemo na izvestan, konačan broj decimala, onda možemo da prihvatimo prethodnu napomenu.

2. TEORIJSKE OSNOVE

U teorijskoj statistici se problem ključnih podataka odavno razmatra. Rešenje daju tzv. **dovoljne statistike**.

Ukratko ćemo navesti definiciju i postupak određivanja dovoljnih statistika i navesti neke primere.

Dovoljna statistika za neki parametar u raspodeli obeležja je ona funkcija uzorka koja sadrži sve informacije o posmatranom parametru. Dovoljna statistike se definiše preko uslovne raspodele, a u primenama se, za konkretno određivanje dovoljnih statistika ne koristi definicija, već sledeća, Fišer-Nojmanova, teorema o faktorizaciji, prema [1].

Teorema o faktorizaciji

Neka je dat prost slučajni uzorak $(X_1,...,X_n)$ obima n za obeležje X, tj. neka su slučajne promenljive $X_1,...,X_n$ nezavisne i neka sve imaju istu raspodelu kao posmatrano obeležje. Neka posmatrano obeležje ima gustinu raspodele $f(x,\theta)$ koja je funkcija nepoznatog parametra θ i neka je $T_{n,\theta}$ jedna statistika za posmatrani uzorak. Statistika $T_{n,\theta}$ je dovoljna za parametar θ ako i samo ako postoje nenegativne funkcije g i h takve da je

$$\prod_{j=1}^{n} f(x_j,\theta) = g(T_{n,\theta})h(x_1,...,x_n)\cdot$$

Primenom ove teoreme dobijamo poznate rezultate: uzoračka sredina je dovoljna statistika za matematičko očekivanje za obeležje sa normalnom raspodelom, a takođe i to da su uzoračka sredina i uzoračka disperzija, uzete kao par, dovoljna statistika za dvodimenzionalni parametar koji predstavlja par koji čine matematičko očekivanje i disperzija za obeležje sa normalnom raspodelom.

Neka je za neko obeležje dat uzorak $(X_1,...,X_n)$ obima *n*. Neka su, zatim, za isto obeležje dobijeni još neki podaci $X_{n+1},...,X_m$. Ako se sada novi podaci pridruže starim, dobijamo objedinjeni uzorak . Ovde ćemo posmatrati objedinjeni uzorak pod pretpostavkom da polazni podaci NISU sačuvani, ali da imamo neke podatke dobijene na osnovu polaznog uzorka. Takav objedinjeni uzorak ćemo označavati $(\underline{X_1,...,X_n}, X_{n+1},...,X_m)$. U oznakama statistika za objedinjeni uzorak ćemo obavezno dodavati *, da bi se bolje uočavalo šta se odnosi na polazni, a šta na objedinjeni uzorak.

Neka je dat prost slučajni uzorak $(X_1,...,X_n)$ obima *n* za posmatrano obeležje. Ako se za taj uzorak izračuna i sačuva uzorački moment reda *r*

$$A_{r,n} = \frac{1}{n}(X_1^r + ... + X_n^r),$$

onda se za objedinjeni uzorak može izračunati uzorački moment reda *r*, bez polaznih podataka, na sledeći način:

$$A^*_{r,m} = \frac{1}{m}(nA_{r,n} + (X_{n+1}^r + ... + X_m^r))\cdot$$

Osim uzoračkih momenata, koriste se i mnoge druge statistike, pa u nastavku teksta ukazujemo na to koji su podaci ključni za izračunavanje statistika za objedinjeni uzorak i za testiranje statističkih hipoteza za objedinjeni uzorak.

3. NEKA PRAKTIČNA UPUTSTVA

Koristeći prethodno date teorijske osnove, razmotrićemo ideju korišćenja «starih» podataka u objedinjenom uzorku sa «novim» podacima. Dakle, imamo neki uzorak obima ***n*** a zatim još nekoliko «novih» podataka o istoj pojavi. Pod tim podrazumevamo da je (na pravilan način) dobijen još jedan uzorak za isto obeležje. Obim novog uzorka u odnosu na obim starog uzorka je manji. Ako su jednaki, ili ako u novom ispitivanju imamo više podataka nego pre, moguće je da ćemo porediti ta dva skupa podataka da vidimo da li je raspodela obeležja ostala ista, ili ćemo jednostavno odbaciti stare podatke i raditi samo sa novima. Veličina novog uzorka nema uticaja na računske postupke o kojima će biti reči. Stoga, nećemo ništa posebno govoriti o obimima ovih uzoraka.

3.1. Računanje statistika za objedinjeni uzorak

U nastavku teksta navodimo neke često korišćene statistike, videti u [3] i neke njihove kombinacije i pokazujemo kako se, ukoliko je moguće, vrednosti odgovarajućih statistika mogu dobiti za objedinjeni uzorak.

(1) Uzoračka sredina

Ako je za prost slučajni uzorak $(X_1,...,X_n)$ obima *n* izračunata uzoračka sredina (srednja vrednost) $\overline{X}_n$, i ako su jedino sačuvane vrednosti $\overline{X}_n$ i *n*, onda se za objedinjeni uzorak $(\underline{X_1,...,X_n}, X_{n+1},...,X_m)$, u kome je prvih *n* podataka izostavljeno, uzoračka sredina računa po formuli

$$\overline{X}^*_m = \frac{1}{m}(n\overline{X}_n + \sum_{k=n+1}^{m} X_k)\cdot$$

Formula sledi neposredno iz definicije uzoračke sredine.

(2) Uzoračka sredina i uzoračka disperzija

Ako su za prost slučajni uzorak obima *n* izračunate i sačuvane uzoračka sredina $\overline{X}_n$ i uzoračka disperzija $\overline{S}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$, onda se uzoračka disperzija za objedinjeni uzorak računa po formuli

$$\overline{S}^{2*}_m = \frac{1}{m}\left(n(\overline{S}_n^2 + (\overline{X}_n)^2) + \sum_{k=n+1}^{m} X_k^2\right) - (\overline{X}^*_m)^2$$

Dokaz sledi iz poznate formule

$$\overline{S}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - (\overline{X}_n)^2,$$

a uzoračka sredina za objedinjeni uzorak se računa po formuli iz (1).

(3) Uzoračka sredina i uzorački koeficijent varijacije

Ako su za prost slučajni uzorak obima *n* izračunate i sačuvane srednja vrednost (uzoračka sredina) $\overline{X}_n$ i uzorački koeficijent varijacije CV_n, onda se uzorački koeficijent varijacije za objedinjeni uzorak računa po formuli

$$CV_m^* = \frac{\sqrt{\frac{1}{m}\left[n(CV_n^2+1)(\overline{X}_n)^2 + \sum_{k=n+1}^{m} X_k^2\right] - (\overline{X}_m^*)^2}}{\overline{X}_m^*}$$

Dokaz proizilazi iz formule

$$CV_n = \frac{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2}}{\overline{X}_n} = \frac{\sqrt{\frac{1}{n}\sum_{j=1}^{n} X_j^2 - (\overline{X}_n)^2}}{\overline{X}_n},$$

a uzoračka sredina za objedinjeni uzorak se računa po formuli iz (1).

(4) Uzoračka sredina, uzorački koeficijent varijacije i uzorački koeficijent asimetrije

Ako su za prost slučajni uzorak obima *n* izračunate i sačuvane veličine: srednja vrednost (uzoračka sredina) $\overline{X}_n$, uzorački koeficijent varijacije CV_n i koeficijent asimetrije (prvi Pirsonov koeficijent)

$$\pi_{1,n} = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^3}{\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2}\right)^3},$$

onda se uzorački koeficijent asimetrije za objedinjeni uzorak računa po formuli

$$\pi_{1,m}^* = \frac{A_{3,m}^* - 3\overline{X}_m^* \cdot A_{2,m}^* + 2(\overline{X}_m^*)^3}{\left[CV_m^* \cdot \overline{X}_m^*\right]^3},$$

gde je

$$A_{3,m}^* = \frac{1}{m}(nA_{3,n} + \sum_{k=n+1}^{m} X_k^3), \qquad A_{2,m}^* = \frac{1}{m}(nA_{2,n} + \sum_{k=n+1}^{m} X_k^2),$$

dok je

$$A_{3,n} = (\overline{X}_n)^3(\pi_{1,n} \cdot CV_n^3 - 2) + 3\overline{X}_n \cdot A_{2,n}$$

i $A_{2,n} = (\overline{X}_n)^2(CV_n^2 + 1)$.

Uzoračka sredina i koeficijent varijacije za objedinjeni uzorak se računaju prema prethodno datim formulama.

(5) Uzoračka sredina, uzorački koeficijent varijacije, uzorački koeficijent asimetrije i uzorački koeficijent spljoštenosti

Ako su za prost slučajni uzorak obima *n* izračunate srednja vrednost (uzoračka sredina) $\overline{X}_n$, uzorački koeficijent varijacije CV_n, koeficijent asimetrije $\pi_{1,n}$ i koeficijent spljoštenosti (drugi Pirsonov koeficijent)

$$\pi_{2,n} = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^4}{\left(\sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2}\right)^4} - 3,$$

onda se uzorački koeficijent spljoštenosti za objedinjeni uzorak računa po formuli

$$\pi_{2,m}^* = \frac{A_{4,m}^* - 4\overline{X}_m^* \cdot A_{3,m}^* + 6(\overline{X}_m^*)^2 \cdot A_{2,m}^* - 3(\overline{X}_m^*)^4}{\left[CV_m^* \cdot \overline{X}_m^*\right]^4} - 3$$

gde je

$$A_{4,m}^* = \frac{1}{m}(nA_{4,n} + \sum_{k=n+1}^{m} X_k^4),$$

$$A_{4.n} = (\overline{X}_n)^4((3 + \pi_{3,n}) \cdot CV_n^4 + 3) + 4\overline{X}_n \cdot A_{3,n} - 6(\overline{X}_n)^2 \cdot A_{2,n}$$

dok se A_3, A_2, A_3^*, A_2^* računaju po formulama datim u okviru (4).

(6) Uzoračka medijana

Uzoračka medijana se ne može odrediti za objedinjeni uzorak na osnovu vrednosti uzoračke medijane za polazni uzorak. Razlog je u činjenici da se medijana računa na osnovu celog varijacionog niza (tj. svih podataka iz uzorka poređanih u neopadajući niz), pa prema tome bez svih podataka ne može da se rekonstruiše.

(7) Raspon uzorka

Ako je dat samo raspon uzorka, onda se ne može rekonstruisati raspon za objedinjeni uzorak, ali ako su sačuvane minimalna $X_{\min}$ i maksimalna $X_{\max}$ vrednost polaznog uzorka, onda se raspon objedinjenog uzorka računa po formuli

$$R^* = X^*_{\max} - X^*_{\min} ,$$

gde je

$$X^*_{\max} = \max\{X_{\max}, X_{n+1},..., X_m\},$$
$$X^*_m = \min\{X_{\min}, X_{n+1},..., X_m\}.$$

(8) Uzoračke sredine, disperzije i koeficijent korelacije

Ako su za proste slučajne uzorke $(X_1,...,X_n)$ i $(Y_1,...,Y_n)$ obima n izračunate srednje vrednosti (uzoračke sredine) $\overline{X}_n$ i $\overline{Y}_n$, uzoračke disperzije $\overline{S}^2_{nX}$ i $\overline{S}^2_{nX}$ i uzorački koeficijent korelacije

$$\rho = \frac{\frac{1}{n}\sum_{j=1}^{n} X_j Y_j - \overline{X}_n \overline{Y}_n}{\sqrt{\overline{S}^2_{nX}\overline{S}^2_{nY}}}$$

onda se uzorački koeficijent korelacije za objedinjeni uzorak

$$(\underline{X_1,...,X_n}, X_{n+1},..., X_m),$$
$$(\underline{Y_1,...,Y_n}, Y_{n+1},..., Y_m)$$

računa po formuli

$$\rho^* = \frac{A^*_{mXY} - \overline{X}^*_m \overline{Y}^*_m}{\sqrt{\overline{S}^{2*}_{mX}\overline{S}^{2*}_{mY}}},$$

gde je

$$A^*_{mXY} = \frac{1}{m}(nA_{nXY} + \sum_{k=n+1}^{m} X_k Y_k),$$
$$A_{nXY} = \rho\sqrt{\overline{S}^2_{nX}\overline{S}^2_{nY}} + \overline{X}_n \overline{Y}_n ,$$

a uzoračke sredine i disperzije za objedinjeni uzorak se računaju po formulama datim pod (1) i (2).

ZAKLJUČAK

Može da se zaključi da se za objedinjeni uzorak mogu rekonstruisati statistike koje sadrže samo uzoračke momente, i to samo na osnovu vrednosti odgovarajućih uzoračkih momenata za polazni uzorak, i vrednosti obima polaznog uzorka. Na taj način, ako na početku istraživanja i nemamo neke specifične zahteve, možemo da sačuvamo vrednosti uzoračkih momenata polaznog uzorka i njegov obim, pa na osnovu tih podataka, kad se kasnije ukaže potreba, možemo da rekonstruišemo vrednosti uzoračkih momenata za objedinjeni uzorak i tako dobijemo vrednosti potrebnih statistika za objedinjeni uzorak.

3.2. O statističkim testovima koji se mogu sprovesti na objedinjenom uzorku na osnovu rezultata tih istih testova na polaznom uzorku

Pri proučavanju slučajnih pojava statističkim metodama često postavljamo hipoteze o parametrima u raspodeli obeležja ili o samoj raspodeli obeležja. Te hipoteze proveravamo statističkim testovima na osnovu uzorka. Ako se za polazni uzorak sačuvaju neki podaci, onda se izvesni testovi mogu sprovesti i na objedinjenom uzorku, iako više nemamo sačuvane polazne podatke.

a) Parametarski testovi koji se odnose na matematičko očekivanje i disperziju se mogu rekonstruisati. Treba sačuvati uzoračku sredinu i disperziju za polazni uzorak, a uzoračka sredina i disperzija za objedinjeni uzorak se računaju kao što je prethodno navedeno pod (1) i (2). Uopšte, prema zaključku datom u vezi statistika koje se mogu rekonstruisati, parametarski testovi, u kojima se koriste statistike zasnovane na uzoračkim momentima, se mogu rekonstruisati.

b) Neparametarski testovi
Kod neparametarskih testova imamo različite slučajeve. Navešćemo samo nekoliko testova.

(1) χ^2 test se može rekonstruisati ako se sačuvaju granice klasa i empirijske frekvencije po klasama. To je dovoljno da se izračuna vrednost test-statistike na objedinjenom uzorku. Ukoliko se, pak, ocenjuju i neki parametri u raspodeli obeležja, onda bismo njihove vrednosti na objedinjenom uzorku izračunali, ukoliko je to moguće, po nekoj od prethodno datih formula.

(2) Test Kolmogorova-Smirnova se ne može rekonstruisati ako nemamo sve polazne podatke, s

obzirom da za empirijsku funkciju raspodele treba daznamo varijacioni niz za polazni uzorak.

(3) Test tačaka zaokreta, jedan od testova za proveru slučajnosti, videti u [2], može se rekonstruisati ako imamo sačuvanu vrednost test statistike, ukupan broj tačaka zaokreta i poslednja dva podatka iz polaznog uzorka, jer onda možemo da odredimo i ukupan broj tačaka zaokreta za objedinjeni uzorak.

(4) Test tačaka rasta (videti u [2]) se može rekonstruisati ako imamo vrednost test-statistike, ukupan broj tačaka rasta i poslednji podatak polaznog uzorka, jer onda možemo da odredimo i ukupan broj tačaka rasta za objedinjeni uzorak.

LITERATURA

[1] Stojanović S., Matematička statistika, Beograd, Naučna knjiga, 1980.

[2] Mališić J., Analiza vremenskih serija, Beograd, Matematički fakultet, 2002.

[3] Jevremović V., Mališić J., Statističke metode u meteorologiji i inženjerstvu, Beograd, Savezni hidrometeorološki zavod, 2002

DATA REDUCTION IN STATISTICS

by

Vesna JEVREMOVIĆ, Jovan MALIŠIĆ

Summary

In this paper we investigate data reduction in statistics. The problem we deal with arises when we want to calculate statistics, or run statistical tests but instead of the whole sample we only have the newest data. This problem could be solved if we have calculated values for some basic statistics for the missing set of data.

Key words: sample moments, coefficient of variation, coefficient of assymetry, coefficient of excess, median, tourning points test, the Kolmogoroff-Smirnoff test, χ^2-test

Redigovano 22.06.2004.