Verovatnoća i statistika – idealni model i pojavni oblici

Dr Biljana Popović, redovni profesor
Prirodno–matematički fakultet u Nišu

3. april 2004. godine

Matematička statistika je primenjena matematička disciplina srodna teoriji verovatnoće. Bazira se na pitanjima i metodima teorije verovatnoće, ali rešava svoje specifične (probleme) zadatke svojim metodama. (Svaka matematička teorija se razvija u okviru nekog modela koji opisuje odredjeni krug realnih pojava čijim se proučavanjem i bavi data teorija.)

U teoriji verovatnoće se polazi od pretpostavke da je poznat prostor verovatnoća $(\Omega, \mathcal{F}, P)$, gde je Ω skup svih elementarnih ishoda, $\mathcal{F}$ je σ-algebra na skupu Ω a P je verovatnoća.

Verovatnoća P, u praktičnim problemima koje treba rešavati, nije u potpunosti poznata. U većini slučajeva se pretpostavlja da $P \in \mathcal{P}$, gde je $\mathcal{P} = \{P\}$ familija verovatnoća. Takvi praktični problemi nazivaju se statističkim modelima.

Dakle, za razliku od modela teorije verovatnoća, statistički model je $(\Omega, \mathcal{F}, \mathcal{P})$.

Primer 1. (Šema Bernulija.) Obavlja se n nezavisnih opita u kojima se realizuje 0 ili 1 sa verovatnoćama redom $1 - p = q$ i p, $0 \leq p \leq 1$. Ishod ovog eksperimenta je

$$\Omega = \{\omega : \omega = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n), \varepsilon_i = 0, 1\}.$$

Pri tome je verovatnoća pojedinog elementarnog ishoda

$$P(\omega) = p^{\sum \varepsilon_i} q^{n - \sum \varepsilon_i}.$$

Ako verovatnoća p nije prethodno poznata, označićemo je sa θ i tu oznaku ćemo nadalje koristiti za svaki nepoznati parametar. U tom slučaju jedina informacija koju imamo o parametru ovog primera je da je $\theta \in \Theta = [0, 1]$. Tačnije, imamo jedino informaciju da raspodela verovatnoća kojom ovaj eksperiment opisujemo pripada familiji $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, gde je $P_\theta = \theta^{\sum \varepsilon_i}(1 - \theta)^{n - \sum \varepsilon_i}$. $\triangle$

U prethodnom primeru je definisan jedan statistički model, dakle model koji u sebi sadrži neku vrstu neodredjenosti. Zadatak matematičke statistike je da se korišćenjem informacije dobijene posmatranjem ishoda eksperimenta, dakle statističkih podataka, smanji ta neodredjenost, odnosno da se, što je moguće tačnije, izvrši izbor $P \in \mathcal{P}$.

Matematička statistika je nauka o statističkom zaključivanju. Statističko zaključivanje podrazumeva rešavanje zadataka obrnutih od onih koje rešava teorija verovatnoće: ona utvrdjuje strukturu statističkih modela prema rezultatima sprovedenih posmatranja, dakle, odredjuje prostor verovatnoća na osnovu eksperimenta. Pri tome posmatranja ne mogu biti proizvoljna. Naime, ona moraju biti ekvivalentna statističkom eksperimentu:

- može se ponavljati proizvoljan broj puta pod istim uslovima,
- unapred je definisano šta se registruje u eksperimentu pri čemu su poznati svi mogući ishodi i
- ishod pojedinačnog eksperimenta nije unapred poznat.

Za prve svesne pokušaje definisanja i primene statističkog zaključivanja uzimaju se popisi stanovništva koje su sprovodili vladari još nekoliko vekova pre naše ere radi utvrdjivanja broja vojnih podanika ili poreskih obveznika. Zasnivanje statistike kao nauke vezuje se za pojavu škole "političkih aritmetičara" u Engleskoj u $XVII$ veku. Po nekima, delo "Natural and Political Observations upon the Bills of Mortality", koje je napisao Dž. Grant (J. Graunt) i objavio 1622. godine, označava početak statistike kao nauke. Dugo vremena je statistika smatrana naučnim metodom za proučavanje društvenih nauka. Medjutim, matematičari koji su neminovno bili uključeni u konstituisanje, formalno definisanje, i postali odgovorni za razvoj statističkog metoda zaključivanja, odgovorni su i za početak primene statistike u prirodnim naukama. Tu ideju medju prvima je prihvatio engleski biolog Galton (Sir Francis Galton, 1822-1911), koji je primenio statistički metod u istraživanjima u biologiji. Teorijski doprinos razvoju matematičke statistike dao je medju prvima švajcarski matematičar Jakob Bernuli (Jacob Bernoulli, 1654-1705) definišući i obrazlažući zakon velikih brojeva u svom delu "Ars conjectandi". Krupan korak u tom pravcu dao je i francuski astronom i matematičar Laplas (Pierre Simon, Marquis de Laplace, 1749-1827). Poznato je njegovo delo "Théorie analytique de probabilités". Buran razvoj matematičke statistike kao teorijske discipline u XX veku omogućen je, pre svega, razvojem teorije verovatnoća u ovom periodu.

1 Osnovni pojmovi statistike

Statistički eksperiment se izvodi nad elementima nekog skupa na kojima se posmatra jedno ili više zajedničkih svojstava.

Definicija 1. *Populacija* ili *generalni skup* je skup elemenata čija se zajednička svojstva izučavaju statističkim metodima. Populacija se simbolički beleži sa Ω, a njen element sa ω.

Definicija 2. *Obeležje* je zajedničko svojstvo elemenata jedne populacije (koje se ispituje). Obeležje može biti kvantitativno (numeričko) ili kvalitativno (atributivno).

Pri izvodjenju statističkog eksperimenta polazi se od pretpostavke da se tom prilikom realizuju neki slučajni dogadjaji. Dakle, pretpostavlja se da se ishod eksperimenta može

prikazati slučajnom veličinom X. Ukoliko je eksperiment ponavljan n puta, ishod se predstavlja slučajnim vektorom $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Pri proučavanju ovog slučajnog vektora poželjno je poznavati njegovu raspodelu. S tim u vezi reći ćemo da treba odrediti gustinu raspodele obeležja, a nadalje ćemo to pojasniti. Ovde će se koristiti termin gustina raspodele u uopštenom značenju, tj. vezivaće se i za slučajne promenljive diskretnog tipa.

Primer 2. Za slučajnu promenljivu sa binomnom raspodelom $\mathcal{B}(1, p)$, kazaćemo da ima gustinu raspodele

$$f(x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & x \neq 0, 1 \end{cases} \quad . \quad \triangle$$

Neka je Y slučajna promenljiva definisana kao funkcija slučajnih promenljivih

$$X_1, X_2, \ldots, X_n,$$

tj. neka je $Y = u(X_1, X_2, \ldots, X_n)$. Odredjivanje gustine raspodele ove slučajne promenljive na osnovu poznavanja zajedničke gustine raspodele vektora slučajnih promenljivih $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, u oznaci $f(x_1, x_2, \ldots, x_n)$, $(x_1, x_2, \ldots, x_n) \in R^n$, je jedan od zadataka matematičke statistike. Sam slučajni vektor $\mathbf{X}$ i funkcije od njegovih komponenata su okosnica matematičke statistike.

Definicija 3. *Uzorak* je deo populacije na kome se ispituje posmatrano obeležje. Broj elemenata u uzorku se naziva *obim uzorka*.

Na uzorku se sprovodi statistički eksperiment. Ishod tog eksperimenta će biti vektor $\mathbf{X}$, koji je po svojim karakteristikama slučajna promenljiva.Vektor $\mathbf{X}$ još zovemo *slučajnim uzorkom* za razliku od njegove *realizovane vrednosti* po obavljenom eksperimentu.

Definicija 4. Vektor $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ koji predstavlja realizaciju vektora $\mathbf{X}$ po obavljenom eksperimentu zovemo *realizovani uzorak*.

U daljem tekstu će se pod uzorkom podrazumevati slučajni uzorak, a kada bude reči o realizovanom uzorku, to će biti naglašeno.

Detaljnije o uzorku i načinima za izbor uzoraka pripada posebnoj oblasti matematičke statistike koja se zove *Teorija uzoraka*.

2 Slučajna promenljiva i obeležje

Populacija ima nešto širi smisao od izvesnog dogadjaja u teoriji verovatnoće, dok je obeležje nešto širi pojam od pojma slučajne promenljive. Naime, izvesan dogadjaj je skup svih mogućih elementarnih ishoda jednog eksperimenta, pri čemu se podrazumevaju različiti ishodi. Populacija je, medjutim, skup svih elemenata na kojima se posmatra neko svojstvo (skup ljudi, skup sijalica, deo tla, itd.). Obeležje je funkcija iz skupa Ω, populacije, u skup koji čine kategorije jednog svojstva. Preciznije, na skupu Ω se definiše relacija ekvivalencije: "dva elementa populacije su u relaciji ako su im jednake vrednosti obeležja koje se na elementima populacije posmatra". Tom relacijom se vrši razbijanje

skupa Ω na klase ekvivalencije, odnosno, definiše se faktor skup. Klase ekvivalencije su kategorije, te se najpre definiše preslikavanje populacije na faktor skup tako što se svakom elementu populacije pridružuje njegova klasa ekvivalencije. Iz faktor skupa je moguće definisati novu funkciju sa vrednostima u skupu realnih brojeva, R, koja je, zapravo, slučajna promenljiva, a u žargonu matematičke statistike, kaže se da se ovom funkcijom vrši kodiranje vrednosti obeležja. U tom smislu se može govoriti o raspodeli obeležja posredstvom raspodele ovako definisane slučajne promenljive, te će se i obeležje, kao i slučajna promenljiva, označavati velikim slovom latinice sa kraja abecede, X,Y,Z,.... U vezi sa uopštenjem pojma gustine raspodele smatraće se da svako obeležje ima svoju gustinu raspodele.

Primer 3. Za populaciju ćemo uzeti studente Prirodno-matematičkog fakulteta u Nišu. Neka je obeležje koje posmatramo na toj populaciji "obrazovni profil". U ovom momentu ćemo posmatrati samo osnovni profil, tj. matematika, fizika, hemija, biologija, geografija. Ovih 5 kategorija bi činile razbijanje skupa Ω. Dakle, studenti istog odseka – obrazovnog profila bi činili jednu klasu ekvivalencije. Nadalje bismo svakom odseku pridružili broj (kod), recimo neka su to prirodni brojevi od 1 do 5. Time bi bila definisana slučajna promenljiva. $\triangle$

Sa gledišta matematičke statistike dato obeležje X je potpuno odredjeno ako je odredjena njegova raspodela, $P\{X \in S\}$, gde je $S \in \mathcal{B}_1$, a $(R, \mathcal{B}_1, P)$ fazni prostor. To je istovremeno i jedan od glavnih problema kojima se bavi matematička statistika: odredjivanje raspodele obeležja. Pri tome je moguće da unapred nije poznata familija dopustivih raspodela ili da je ona poznata, a da iz nje treba napraviti pravi izbor ocenom vrednosti nepoznatih parametara koji u raspodeli figurišu. Dakle, osnovni problem statističkog zaključivanja je da na osnovu statističkog eksperimenta nešto zaključi o raspodeli obeležja.

3 Zaključivanje na osnovu uzorka

Zaključivanje o raspodeli obeležja vrši se na osnovu izabranog uzorka. Otuda je važno da izabrani uzorak bude reprezentativan, tj. da bude takav da se sa dovoljnom tačnošću zaključak o raspodeli posmatranog obeležja dobijenoj na uzorku može da ekstrapoluje na čitavu populaciju.

Okosnica naučne oblasti koju zovemo matematičkom statistikom ili, jednostavno, statistikom, je funkcija od uzorka koja je osnovni alat u procesu statističkog zaključivanja, a koja je opisana sledećom definicijom:

Definicija 5. *Statistika* je funkcija od uzorka čiji analitički izraz ne zavisi od nepoznatih parametara obeležja, tj. funkcija od uzorka i poznatih konstanata.

Primeri nekih statistika su:

$$T_n = \sum_{i=1}^{n} X_i \quad - \text{total} \quad \text{uzorka}$$

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad - \text{sredina uzorka}$$

$$\overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad - \text{disperzija uzorka}$$

$$\overline{S}_n = \sqrt{\overline{S}_n^2} \quad - \text{uzoračka standardna devijacija}$$

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad - \text{popravljena disperzija uzorka}$$

$$R = X_{max} - X_{min} \quad - \text{raspon uzorka}\,.$$

Za dva obeležja X i Y i uzorak $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ iz populacije na kojoj se posmatra dvodimenziono obeležje (X, Y) može se definisati statistika

$$R_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\overline{S}_X \overline{S}_Y} \quad - \text{uzoračkI koeficijent korelacije}\,,$$

gde su sa $\overline{S}_X$ i $\overline{S}_Y$ označene uzoračke standardne devijacije za obeležja X i Y redom.

Posebno mesto medju statistikama imaju tzv. statistike poretka. Ove se statistike definišu posredstvom varijacionog niza:

Definicija 6. *Varijacioni niz* čine elementi uzorka poredjani u neopadajućem poretku.

Za uzorak $(X_1, X_2, \ldots, X_n)$ varijacioni niz čini niz slučajnih promenljivih sačinjen od elemenata ovog uzorka u oznaci $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ za koji važi

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \quad .$$

Za realizovane vrednosti varijacionog niza koristi se isti termin *varijacioni niz*, bez opasnosti od zabune, a označavaju se malim slovima:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \quad .$$

Definicija 7. *Statistika poretka reda k* uzorka obima n, $1 \leq k \leq n$, je k–ti element varijacionog niza posmatranog uzorka, dakle slučajna promenljiva $X_{(k)}$.

Neka je uzorak $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ prost slučajni uzorak iz populacije sa obeležjem X čija je funkcija raspodele F. **U definisanju funkcije raspodele biće sve vreme korišćena neprekidnost s desna.** Za svako $x \in R$ definisaćemo slučajnu veličinu $\mu_n(x)$ kao broj elemenata uzorka $\mathbf{X}$ koji su manji ili jednaki x, tj.

Definicija 8.

$$\mu_n(x) = card\{j | X_j \leq x, j = 1, 2, \ldots, n\} \quad , \quad x \in R \quad .$$

Nadalje se može definisati slučajna promenljiva $S_n(x)$ koja daje vrednosti slučajne promenljive $\mu_n(x)$ u relativnom odnosu prema obimu uzorka:

DEFINICIJA 9. *Empirijska funkcija raspodele* uzorka $\mathbf{X}$ je statistika

$$S_n(x) \stackrel{\text{def}}{=} \frac{\mu_n(x)}{n} \quad , \quad x \in R \quad .$$

Slučajna promenljiva $S_n(x)$ je statistika čiji je kodomen skup

$$\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$$

ili njegov pravi podskup sa verovatnoćama

$$P\{S_n(x) = k/n\} = P\{\mu_n(x) = k\} = \binom{n}{k}(F(x))^k(1-F(x))^{n-k}, \quad k = 0, 1, \ldots, n.$$

Ovo otuda što, prema definiciji, slučajna promenljiva $\mu_n(x)$ ima binomnu raspodelu, $\mathcal{B}(n,p)$ sa $p = P\{X \leq x\} = F(x)$, $x \in R$. Statistiku $S_n(x)$ možemo posmatrati i kao aritmetičku sredinu indikatora

$$I_{A_i} = \begin{cases} 1, & \omega \in A_i \\ 0, & \omega \notin A_i \end{cases} ,$$

$A_i = \{\omega | X_i(\omega) \leq x\}$, a s obzirom da je $E(I_{A_i}) = F(x)$ za fiksirano $x \in R$, važi teorema:

Teorema 1. *Za fiksirano $x \in R$, $S_n(x) \longrightarrow F(x)$, $n \to \infty$ skoro izvesno, tj.*

$$P\{S_n(x) \to F(x)\,,\, n \to \infty\} = 1.\triangle$$

Za realizovani uzorak $(x_1, x_2, \ldots, x_n)$, $S_n(x)$, $x \in R$, je monotono neopadajuća funkcija sa mogućim skokovima u tačkama varijacionog niza $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$:

$$S_n(x) = \frac{k}{n}, \quad x \in [x_{(k)}, x_{(k+1)}), k = 0, 1, \ldots, n \quad .$$

Pri tome su uvedene oznake $x_{(0)} = -\infty$, i u tom slučaju je i leva granica intervala otvorena, i $x_{(n+1)} = +\infty$. Ukoliko su svi elementi u realizovanom uzorku različiti, skokovi su veličine $1/n$.

Konvergencija o kojoj je bilo reči u prethodnoj teoremi, ostvaruje se i uniformno po $x \in R$. O tome govori tzv. **centralna teorema matematičke statistike**. Jedan od njenih oblika je sledeći.

Teorema 2 (Glivenko-Kanteli) *Neka je F funkcija raspodele obeležja X i $S_n(x)$, $x \in R$, empirijska funkcija raspodele uzorka obima n iz populacije sa obeležjem X. Tada važi*

$$P\{\sup_{x \in R} |S_n(x) - F(x)| \to 0\,,\, n \to \infty\} = 1.\triangle$$