Matematika i statistika

Doktorski studij Strojarskog fakulteta

Sveučilište u Slavonskom Brodu

Predavanje 1

Predavač:
izv.prof.dr.sc. Nenad Šuvak, Odjel za matematiku, Sveučilište u Osijeku

SADRŽAJ KOLEGIJA

bavit ćemo se sljedećom problematikom:

- kreiranje uzorka i prikupljanje podataka
- deskriptivna (opisna) statistika
- statistička analiza
  - procjena nepoznatih parametara
  - testiranje statističkih hipoteza
  - statističko modeliranje
- programska platforma za statističke procedure - R

STATISTIKA

- **statistika** - znanstvena disciplina koja se bavi razvojem metoda **prikupljanja**, **opisivanja** i **analiziranja** podataka te primjenom tih metoda u procesu donošenja zaključaka na temelju prikupljenih podataka
- **statističko istraživanje** - fokusirano na skup **objekata**, tj. **jedinki** (ljudi, životinja, biljaka, stvari, država, gradova, poduzeća, itd.) i skup odabranih veličina (**varijabli**) koje se na njima promatraju

Prikupljanje i organizacija podataka

Populacija i uzorak

- **populacija** - skup svih jedinki koje su predmet istraživanja (zadovoljavaju svojstva bitna za obilježje koje se istražuje)
- **uzorak** - podskup populacije na kojemu je osigurano kvalitetno provođenje istraživanja
- **reprezentativan uzorak** - uzorak u kojem su zastupljene sve tipične karakteristike populacije bitne za istraživanje
- **slučajni uzorak** - uzorak konstruiran u skladu sa zahtjevom da svaka jedinka iz populacije ima istu šansu biti izabrana u uzorak
- **varijabla** - obilježje koje se istražuje
- **baza podataka** - tablica koja za svaku jedinku iz uzorka sadrži vrijednosti svih varijabli obuhvaćenih istraživanjem (redovi tablice - jedinke; stupci tablice - varijable)

PRIMJER

Baza podataka **djelatnici.csv** sadrži podatke o uzorku djelatnika neke tvornice:

- varijabla **spol** sadrži informaciju o spolu (**M** - muški spol, **Z** - ženski spol)
- varijabla **odjel** sadrži naziv odjela u kojem je djelatnik zaposlen (**TR** - transport, **P**- pakiranje, **IS** - isporuka)
- varijabla **obrazovanje** sadrži stručnu spremu djelatnika (SSS - srednja stručna sprema, VŠSS - viša stručna sprema, VSS - visoka stručna sprema)
- varijabla **dob** sadrži starost djelatnika u godinama
- varijabla **visina** sadrži visinu djelatnika u centimetrima

PRIMJER

- varijabla **rukovostvo** sadrži broj godina rada koje je djelatnik proveo na nekoj od rukovodećih pozicija u toj tvornici
- varijabla **placa_prije** sadrži iznos godišnje plaće djelatnika prije reorganizacije poslovnog sustava
- varijabla **placa_poslije** sadrži iznos godišnje plaće djelatnika nakon reorganizacije poslovnog sustava
- varijabla **placa_konkurencija** sadrži iznos godišnje plaće djelatnika na istom radnom mjestu u konkurentskoj tvornici.

Opisivanje varijabli - korištenjem procedura u **R-u**

KVALITATIVNE VARIJABLE

kvalitativne varijable - varijable čije vrijednosti nisu realni brojevi nego njihove vrijednosti svrstavamo u kategorije, npr.

- spol (m ili ž)
- krvne grupe (A, B, AB, 0)
- boja očiju (plava, smeđa, zelena)
- opisne ocjene (ništa, malo, srednje, puno)
- radna mjesta u nekom poduzeću (pripravnik, stručni suradnik, viši stručni suradnik, . . .)

KVALITATIVNE VARIJABLE

kvalitativne varijable - mogu biti nominalne i ordinalne

- nominalne - vrijednosti su kategorije među kojima se ne može uspostaviti prirodan poredak (npr. spol i krvne grupe)
- ordinalne - među kategorijama se može uspostaviti prirodan poredak (npr. stručna sprema i numeričke ocjene u školi)

NUMERIČKE VARIJABLE

- **numeričke varijable** - varijable čije su vrijednosti realni brojevi, npr.
  - starost osobe
  - temperatura mora
  - koncentracija soli u morskoj vodi
  - vodostaj rijeke
  - broj neispravnih prizvoda na proizvodnoj liniji
  - broj bodova na testu
  - postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine
  - visina odobrenog kredita u nekoj valuti
- kategorije kvalitativnih varijabli mogu se izražavati brojevima, no to ih ne čini numeričkim varijablama
- među numeričkim varijablama razlikujemo **diskretne** i **neprekidne** varijable

DISKRETNE NUMERIČKE VARIJABLE

diskretne numeričke varijable - mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti, npr.

- broj neispravnih prizvoda na proizvodnoj liniji
- broj bodova na testu

NEPREKIDNE NUMERIČKE VARIJABLE

- **neprekidne numeričke varijable** - skup mogućih vrijednosti je cijeli skup realnih brojeva ili neki interval, npr.
  - starost osobe
  - temperatura mora
  - koncentracija soli u morskoj vodi
  - vodostaj rijeke
  - postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine
  - visina odobrenog kredita u nekoj valuti
- u svrhu prikaza podataka i nekih statističkih analiza, vrijednosti numeričke varijable se kategoriziraju

R - STATISTIČKI PROGRAMSKI JEZIK

- **instalacijski paket** - http://www.r-project.org
- **praktičniji R editor** - https://www.rstudio.com (odabrati RStudio Desktop Open Source Edition)
- **korisničko sučelje RStudia** - podijeljeno na četiri dijela:
  - otvorene skripte za upisivanje koda
  - povijest naredbi i memorija
  - konzola
  - output prozor u kojem se prikazuju datoteke, slike, paketi i pomoć
- **osnovna svojstva R sintakse** - osjetljiv na mala i velika slova; komentari započinju znakom ♯; dozvoljeno je koristiti točku u imenovanju varijabli

R - STATISTIČKI PROGRAMSKI JEZIK

- R je kalkulator - upisivanje jednostavnih računskih naredbi odmah daje rezultat
- za R postoji nekoliko tisuća paketa koji se jednostavno instaliraju i sadrže gotove procedure za brojne statističke metode
- pokretanjem R-a se učitava dvadesetak paketa, a svi ostali se instaliraju i učitavaju po potrebi:
  - instalacija: install.packages('ime.paketa')
  - učitavanje: library('ime.paketa')
- R help - označiti ključnu riječ i stisnuti F1
- za učitavanje datoteka potrebno je postaviti radni direktorij u kojemu se te datoteke nalaze
- preporuka je raditi s datotekama (bazama podataka) u csv formatu

METODE OPISIVANJA KVALITATIVNIH VARIJABLI

- **kvalitativne varijable** - vrijednosti tih varijabli su kategorije
- pri proučavanju kvalitativnih varijabli pažnju usmjeravamo na zastupljenost pojedine kategorije u uzorku
- **frekvencija kategorije, relativna frekvencija kategorije** - mjere kojima opisujemo zastupljenost kategorije u uzorku

Frekvencija kategorije

- neka varijabla (oznaka X) ima k kategorija($k = 4$ znači da varijabla ima 4 kategorije, npr. krvne grupe)
- označimo pojedine kategorije kao $x_1, x_2, \ldots, x_k$
- **frekvencija kategorije x_i, oznaka f_i** - broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i, $i = 1, \ldots, k$
- frekvencija pojedine kategorije ovisi o broju izmjerenih vrijednosti te varijable - poteškoće pri uspoređivanju zastupljenosti različitih kategorija varijable

Relativna frekvencija kategorije

- **relativna frekvencija kategorije** x_i - broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i podijeljen s ukupnim brojem izmjerenih vrijednosti za tu varijablu
- ako je n broj svih izmjerenih vrijednosti promatrane varijable, relativnu frekvenciju kategorije x_i racunamo kao

$$\frac{f_i}{n}$$

- relativna frekvencija kategorije je mjera zastupljenosti koja daje informaciju o udjelu kategorije u uzorku poznate dimenzije i često se izražava kao postotak

Grafički prikazi

frekvencije i relativne frekvencije pojedinih kategorija neke varijable prikazujemo:

- **tablično** - tablicom frekvencija i relativnih frekvencija
- **grafički** - stupčastim i kružnim dijagramima frekvencija i relativnih frekvencija
- u tabličnom i grafičkom prikazu frekvencija i relativnih frekvencija trebaju biti zastupljene sve kategorije varijable
- **R** - primjeri opisa kvalitativnih varijabli iz baze podataka djelatnici.csv

METODE OPISIVANJA DISKRETNIH NUMERIČKIH VARIJABLI

za opis tih varijabli možemo koristiti **iste metode kao pri opisivanju kvalitativnih varijabli:**

- tablice frekvencija i relativnih frekvencija
- grafičke prikaze frekvencija i relativnih frekvencija - stupčaste i kružne dijagrame
- za prikazivanje podataka iz neprekidnih numeričkih varijabli frekvencije, stupčasti i kružni dijagrami napravljeni na osnovu svake pojedine izmjerene vrijednosti su nepregledni i nepraktični

METODE OPISIVANJA NEPREKIDNIH NUMERIČKIH VARIJABLI

razvrstavanje izmjerenih vrijednosti neprekidne numeričke varijable u **kategorije**:

- skup svih izmjerenih vrijednosti razvrstava se u disjunktne intervale
- nema točno definiranog pravila po kojemu bi trebalo definirati duljine intervala niti nihov broj, ali je jasno da ih ne smije biti niti previše niti premalo da bi cijeli postupak imao smisla i služio svrsi
- kriterij za kategorizaciju treba biti temeljen na razumijevanju problema koji proučavamo - podatke ćemo kategorizirati na način koji nam omogućava efikasno dobivanje potrebnih informacija

METODE OPISIVANJA NEPREKIDNIH NUMERIČKIH VARIJABLI - MJERE DESKRIPTIVNE STATISTIKE

karakteristika numeričkih varijabli - među njihovim vrijednostima postoji prirodan **uređaj** i možemo definirati **numeričke karakteristike** tih varijabli koje imaju logičnu interpretaciju i mogu se iskoristiti za njihovo opisivanje:

- mjere centralne tendencije ili mjere sredine
- mjere raspršenosti

MJERE SREDINE - PROSJEK ILI ARITMETIČKA SREDINA

- **aritmetička sredina** (eng. mean) niza podataka $x_1, x_2, \ldots, x_n$ definirana je izrazom

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- za niz podataka $1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8$ (ima ih devet) aritmetička sredina je

$$\frac{1.2 + 2.1 + 3.2 + 4.3 + 5.4 + 6.5 + 7.6 + 8.7 + 9.8}{9} \approx 5.42$$

MJERE SREDINE - MEDIJAN

- niz podataka $x_1, x_2, \ldots, x_n$ sortiramo u uzlaznom poretku - $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$
- **medijan** je mjera centra temeljena na sortiranom nizu podataka - barem pola podataka je manje ili jednako medijanu, a istovremeno je barem pola podataka veće ili jednako od medijana:

$$med(x_1, x_2, \ldots, x_n) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , \quad n \text{ neparan} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & , \quad n \text{ paran} \end{cases}$$

MJERE SREDINE - MEDIJAN

- ako imamo **neparan broj** podataka, onda postoji podatak koji je u sortiranom nizu na srednjoj poziciji - u tom je slučaju taj podatak medijan
- npr. neka je $1, 2, 5, 6, 1, 2, 7, 2, 2, 3, 2$ niz podataka - sortiramo ga u uzlaznom poretku:

$$1, 1, 2, 2, 2, \mathbf{2}, 2, 3, 5, 6, 7$$

- s obzirom da ima 11 podataka, medijan je podatak na šestoj poziciji u sortiranom nizu, tj. 2

MJERE SREDINE - MEDIJAN

- ako imamo **paran broj** podataka, onda ne postoji podatak koji je na srednjoj poziciji jer "srednju poziciju zauzimaju" dva podatka - medijan se tada definira kao polovina između ta dva podatka (tj. aritmetička sredina tih dvaju podataka)
- npr. neka je $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ niz podataka - sortiramo ga u uzlaznom poretku:

$$1, 1, 2, 2, 2, \mathbf{2}, \mathbf{3}, 3, 5, 5, 6, 7$$

- kako niz sadrži 12 podataka, medijan je aritmetička sredina šestog i sedmog podatka, tj. $(2 + 3)/2 = 2.5$

MJERE SREDINE - MOD

- **mod** niza podataka je podatak s najvećom frekvencijom
- mod ne mora biti jedinstven
- u nizu podataka

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$$

vrijednost 2 je izmjerena najviše puta, čak četiri puta, pa je 2 mod ovog niza podataka

- u nizu podataka

$$1, 3, 2, 5, 6, 5, 3, 1, 2, 7, 2, 2, 3, 3$$

vrijednosti 2 i 3 su izmjerene točno četiri puta, pa mod ovog niza podataka nije jedinstven nego su modovi 2 i 3

KVANTILI

- **postotna vrijednost ili kvantil** za neki broj $p \in \langle 0, 100 \rangle$ (oznaka x'_p) je vrijednost za koju je barem $p\%$ podataka manje ili jednako x'_p, a barem $(100 - p)\%$ podataka veće ili jednako x'_p
- dvadesetpet postotna vrijednost - **donji kvartil**
- sedamdesetpet postotna vrijednost - **gornji kvartil**
- kao i kod računanja medijana, ako se u sortiranom nizu na traženoj poziciji za računanje kvantila nalaze dva podatka, određujemo ga kao njihovu aritmetičku sredinu
- razlika gornjeg i donjeg kvartila naziva se **interkvartilni raspon**

KVANTILI

- neka je $1, 2, 5, 6, 6, 1, 3, 7, 3, 3, 3, 3$ niz podataka - prvo ga sortiramo u uzlaznom poretku:

$$1, 1, \mathbf{2}, \mathbf{3}, 3, \mathbf{3}, \mathbf{3}, 3, \mathbf{5}, \mathbf{6}, 6, 7$$

- donji kvartil - kako je od medijana (3) manje ili jednako 6 podataka, donji kvartil je aritmetička sredina podataka na trećoj i četvrtoj poziciji u sortiranom nizu, dakle $(2 + 3)/2 = 2.5$
- gornji kvartil - kako je od medijana veće ili jednako 6 podataka, gornji kvartil je aritmetička sredina podataka na devetoj i desetoj poziciji u sortiranom nizu, dakle $(5 + 6)/2 = 5.5$

MJERE RASPRŠENOSTI - RASPON

- u nizu podataka $x_1, x_2, \ldots, x_n$
  - najmanji nazivamo **(minimum)** - oznaka $x_{\min}$
  - najveći nazivamo **(maksimum)** - oznaka $x_{\max}$
- **raspon** (eng. range) podataka - razlika maksimuma i minimuma
- u nizu podataka $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ minimum je 1, maksimum je 7, a raspon $7 - 1 = 6$

Mjere raspršenosti - maksimalno odstupanje od prosjeka

- **maksimalno odstupanje od prosjeka**, je broj

$$\max\{(\bar{x}_n - x_{\min}), (x_{\max} - \bar{x}_n)\}$$

- za niz podataka $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ je:

$$x_{\min} = 1 \qquad x_{\max} = 7$$

$$\bar{x}_n = \frac{1+2+5+6+5+1+2+7+2+2+3+3}{12} = 3.25$$

$$\max\{3.25 - 1, 7 - 3.25\} = \max\{2.25, 3.75\} = 3.75$$

Mjere raspršenosti - varijanca i standardna devijacija

- **varijanca** i **standardna devijacija** karakteriziraju raspršenost podataka oko aritmetičke sredine
- varijanca podataka $x_1, x_2, \ldots, x_n$:

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- standardna devijacija je kvadratni korijen varijance:

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Mjere raspršenosti - varijanca i standardna devijacija

- za niz podataka $1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8$ aritmetička sredina je približno 5.42, pa su varijanca i standardna devijacija

$$s_n^2 \approx \frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2 \approx 7.87$$

$$s_n \approx \sqrt{\frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2} \approx 2.81$$

Grafički prikazi - kutijasti dijagram (box and whisker plot)

- kutijastim dijagramom prikazujemo odnos pet numeričkih karakteristika niza podataka: minimum, donji kvartil, medijan, gornji kvartil i maksimum
- ako postoje, na kutijastom dijagramu se označavaju i **stršeće vrijednosti**
- **stršeća vrijednost** - podatak koji je značajno veći ili manji u odnosu na druge podatke:
  - ili je netočno izmjeren ili krivo unesen
  - dolazi iz druge populacije
  - točno je izmjeren i unesen u bazu, ali predstavlja rijetku pojavu u populaciji (ekstremna vrijednost)

METODE OPISIVANJA NEPREKIDNIH VARIJABLI - PRIMJER

- određivanje mjera deskriptivne statistike i metode grafičkog opisivanja numeričkih varijabli ilustrirat ćemo u R-u na varijablama **dob**, **placa_prije**, **placa_poslije** i **placa_konkurencija** iz baze podataka **djelatnici.csv**

Opisivanje podataka i statističko zaključivanje

- navedenim metodama samo **opisujemo** podatke prikupljene na uzorku iz populacije
- na temelju podataka želimo donositi argumentirane i jasne **statističke zaključke** na razini populacije
- to znači da **varijable**, koje sadrže podatke prikupljene/izmjerene na uzorku iz populacije, moramo naučiti općenito modelirati - tu se javlja koncept **slučajne varijable**

LITERATURA

- Benšić, M. i Šuvak, N., *Primijenjena statistika*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2013.
- Benšić, M. i Šuvak, N., *Uvod u vjerojatnost i statistiku*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2014.