Osnove statistike u demografiji

Predavanje 7

Procjene parametara jedne populacije

- Zaključci o karakteristikama pojave mogu se donijeti:
  - Prikupljanjem podataka o svim jedinicama promatrane pojave koje čine populaciju (osnovni skup), ili
  - o samo jednom dijelu jedinica
- Uzorak (engl. *sample*)
- Inferencijalna statistika: na temelju uzorka nastoje se donijeti što točniji zaključci o karakteristikama populacije koje se nazivaju parametrima
- Reprezentativan uzorak: dobro opisuje karakteristike populacije (osnovnog skupa)
- Jednostavni slučajni uzorak. Izbor u:
  - konačnim i
  - beskonačnim populacijama

Procjene parametara jedne populacije

- Vrijednosti parametara populacije su nepoznate i one se procjenjuju s pomoću uzorka
- Zaključci o vrijednostima parametara populacije mogu se donijeti samo uz određenu razinu pouzdanosti
- Dvije su skupine metoda inferencijalne statistike:
    - Procjena parametara
    - Testiranje hipoteza

Procjene parametara

- Potrebno je definirati procjenitelja i odrediti oblik njegove distribucije vjerojatnosti (sampling distribucija procjenitelja)
- Mogu se izračunati:
  - procjena parametra jednim brojem i
  - intervalna procjena

Testiranje hipoteza

- Procjenitelju odabranog parametra pridružuje se standardizirana varijabla
- Na temelju podataka iz slučajno izabranog uzorka veličine *n* i poznavanja distribucije vjerojatnosti standardiziranog procjenitelja donosi odluka o odbacivanju ili ne odbacivanju postavljene hipoteze u skladu s odgovarajućim pravilom odlučivanja

Metode izbora uzorka

- Metode inferencijalne statistike temelje se na rezultatima teorije vjerojatnosti.
- Uvjet za primjenu teorije vjerojatnosti je slučajan izbor jedinica u uzorak
- Inferencijalna statistika bavi se poopćavanjem
- Kako bi poopćavanje bilo što kvalitetnije i vjerodostojnije uzorak mora biti reprezentativan
- Reprezentativnost, mogućnost poopćavanja rezultata i primjerenost primjene statističkih metoda ovise o načinu izbora uzorka iz populacije
- Izbor uzorka može biti:
    - namjeran ili
    - slučajan

Metode izbora uzorka

- Namjerni izbor: subjektivan; zato često nije reprezentativan
- Slučajni izbor: objektivan
- Na rezultate slučajnog uzorka može se primijeniti teorija vjerojatnosti, što omogućava poopćavanje:
    - Procjenu parametara
    - Testiranje hipoteza
- Slučajni uzorci nazivaju se još probabilističkima (engl. *random samples*):
    - jednostavni slučajni uzorak,
    - sistematski uzorak,
    - stratificirani uzorak i
    - uzorak skupina
- Plan uzorka je skup pravila i postupaka određenog načina izbora uzorka iz populacije

Jednostavni slučajni uzorak

- Svaka jedinica populacije ima jednaku vjerojatnost da bude izabrana u uzorak

Sistematski uzorak

- Slučajni uzorak u kojem izbor jedinica ovisi o koraku izbora *k* $k = \frac{N}{n}$
- Frakcija izbora $f = \frac{1}{k} = \frac{n}{N}$
- Na primjer, ako se iz populacije *N* = 50 kućanstava u jednom selu izabire uzorak veličine *n* = 10 , tada je korak izbora *k* = 5.
- Popis numeriranih jedinica cijele populacije naziva se okvir izbora

Stratificirani uzorak

- Prikladan kada osnovni skup nije homogen
- Tada se osnovni skup može podijeliti u grupe s obzirom na promatrano obilježje koje se ispituje, tako da jedinice izabrane iz tih grupa budu što homogenije.
- Takve grupe su disjunktni podskupovi i nazivaju se stratumima.
- Stratuma može biti više i ne moraju biti jednakih veličina

Uzorak skupina

- Često se primjenjuje u istraživanju tržišta (segmentacija potrošača)
- Formira se u dva koraka:
  - Cijela se populacija podijeli na konačan broj skupina (klastera) te se na slučajan način izabiru određene skupine (klasteri).
  - Zatim se iz svake izabrane skupine na slučajan način izabiru jedinice

Procjenitelj parametra, vrijednost procjene i sampling-distribucija procjenitelja

- Postupak procjenjivanja nepoznatog parametra populacije provodi se s pomoću procjenitelja
- Procjenitelj (engl. *estimator*) je slučajna varijabla kojom se procjenjuje parametar populacije, dok se konkretna vrijednost procjenitelja, dobivena na uzorku podataka, naziva procjenom (engl. *estimate*).
- "Kvaliteta" procjenitelja ovis o njegovim svojstvima.
- Ako se želi odrediti koja svojstva ima procjenitelj određenog parametra, treba poznavati distribuciju vjerojatnosti tog procjenitelja

Sampling-distribucija procjenitelja

- Konkretne vrijednosti procjenitelja određenog parametara razlikovat će se od uzorka do uzorka.
- Prosječno odstupanje vrijednosti procjenitelja od stvarne vrijednosti parametra populacije kreće se u granicama slučajnih varijacija.
- Te varijacije se nazivaju sampling-varijacije, a distribucija vjerojatnosti procjenitelja se naziva **sampling-distribucija**
- Poželjno je da se vrijednosti procjenitelja na izabranim uzorcima "gomilaju" oko stvarne vrijednosti parametra populacije (nepristranost)
- Da je prosječno odstupanje od stvarne vrijednosti parametra populacije što manje (efikasnost).
- Ponekad je dovoljno da statistička svojstva vrijede asimptotski, tj. da s porastom veličine uzorka *n* procjenitelj ima navedena svojstva

Procjena aritmetičke sredine populacije

Standardna pogreška procjene aritmetičke sredine

- Standardna devijacija procjenitelja $\bar{X}$ je prosječno odstupanje aritmetičkih sredina uzoraka od sredine populacije

$$\sigma_{\overline{X}} = \sqrt{Var(\overline{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

- Kada se slučajni uzorak izabire bez ponavljanja iz konačne populacije, rezultati izbora jedinica u uzorak nisu nezavisni, tj. kovarijance između varijabli uzorka $X_1, X_2, . . . , X_n$ nisu jednake nuli. U tom slučaju je standardna pogreška aritmetičke sredine

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

- Faktor korekcije za konačne populacije $\sqrt{\frac{N-n}{N-1}}$
  - Izostavlja se ako je

$$\frac{n}{N} < 0{,}05$$

Sampling-distribucija procjenitelja $\bar{X}$

- Ako je jednostavni slučajni uzorak izabran iz normalno distribuirane populacije sampling-distribucija procjenitelja $\bar{X}$ je normalna distribucija (za bilo koju veličinu uzorka *n*)

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Ako populacija nije normalno distribuirana, sampling-distribucija procjenitelja $\bar{X}$ bit će približno normalnog oblika samo za velike uzorke ($n > 30$), što je posljedica **centralnog graničnog teorema**

Procjena aritmetičke sredine populacije kada je varijanca populacije poznata

- Intervalna procjena
- Razina pouzdanosti (engl. *confidence level*): $1 - \alpha$

$$P(-z_{\alpha/2} < Z < +z_{\alpha/2}) = P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} < +z_{\alpha/2}\right) = 1 - \alpha.$$

Procjena aritmetičke sredine populacije kada je varijanca populacije poznata

- Interval procjene sredine μ normalno distribuirane populacije s poznatom varijancom σ^2 , uz razinu pouzdanosti $1 - \alpha$

$$P\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Primjer 6.11. Procjena parametra μ jednim brojem i intervalom na temelju velikog uzorka, poznata standardna devijacija populacije

- Bahovec i Erjavec (2015.) str. 262

Procjena aritmetičke sredine populacije kada varijanca populacije nije poznata

$$P(\bar{x} - t_{\alpha/2}\sigma_{\overline{X}} < \mu < \bar{x} + t_{\alpha/2}\sigma_{\overline{X}}) = 1 - \alpha$$

Procjena proporcije populacije

Proporcija populacije	Procjenitelj proporcije	Procjena proporcije jednim brojem
p	$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$	$\hat{p} = \frac{m}{n}$

- Potrebno je utvrditi svojstva procjenitelja od *p*, te odrediti sampling-distribuciju standardiziranog procjenitelja od *p*

Svojstva procjenitelja proporcije populacije

- Pretpostavlja se da je varijabla X Bernoullijeva slučajna varijabla s nepoznatom očekivanom vrijednosti, *p* , koja se procjenjuje na temelju jednostavnog slučajnog uzorka.
- Varijable uzorka X_1 , X_2 ,. . . , X_n su nezavisne i jednako distribuirane.
- Varijable uzorka X_1 , X_2 ,. . . , X_n također su Bernoullijeve slučajne varijable koje su nezavisne, te mogu poprimiti vrijednosti 0 ili 1.
- Procjenitelj parametra p definira se kao aritmetička sredina varijabli uzorka

$$X_i \sim B(1,p)$$

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

Svojstva procjenitelja proporcije populacije

- Zbroj *n* Bernullijevih nezavisnih slučajnih varijabli je slučajna varijabla koja ima binomnu distribuciju

- Očekivana vrijednost i varijanca procjenitelja $\hat{p}$

Svojstva procjenitelja proporcije populacije

- Sampling-distribucija procjenitelja $\hat{p}$ je binomna distribucija
- Procjena proporcije populacije jednim brojem

$$\hat{p} = \frac{m}{n},\ \hat{q} = 1 - \hat{p}.$$

- Standardna pogreška procjene proporcije je $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n-1}}$

- Ako je $\frac{n}{N} \geqslant 0{,}05$ tada je standardna pogreška procjene proporcije

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n-1} \cdot \frac{N-n}{N-1}}.$$

Procjena proporcije populacije s pomoću velikog uzorka

- Proporcija populacije može se procijeniti jednim brojem i intervalom.
- Postupak procjenjivanja proporcije populacije intervalom provodi se samo za velike uzorke.
- Za velike uzorke (n > 30) sampling-distribucija procjenitelja $\hat{p}$ se dobro aproksimira normalnom distribucijom
- Interval procjene proporcije p s pomoću velikog uzorka je

$$P(\hat{p} - z_{\alpha/2}\sigma_{\hat{p}} < p < \hat{p} + z_{\alpha/2}\sigma_{\hat{p}}) = 1 - \alpha$$

Primjer 6.22. Interval procjene proporcije populacije

- Bahovec i Erjavec (2015.) str. 279