Osnove statistike u demografiji

Predavanje 9

Testiranje hipoteza (podsjetnik)

- Jedna populacija
- Jedno obilježje
  - Numeričko (aritmetička sredina)
  - Kvalitativno (proporcija)
- Parametarski testovi: pretpostavka da je distribucija populacije poznata (Normalna distribucija)
  - z-test i *t*-test (test hipoteze o aritmetičkoj sredini)
  - z-test (test hipoteze o proporciji)

Ostali testovi

- Jedna populacija (varijanca)
- Dvije populacije
  - aritmetička sredina (z-test i *t*-test),
    - nezavisni i
    - zavisni uzorci (gleda se ista grupa prije i nakon tretmana)
  - proporcija (veliki uzorci: z-test) i
  - Varijanca (*F*-test)
- Parametarski testovi: pretpostavka da je distribucija populacije poznata (Normalna distribucija)
- Neparametarski testovi (nije poznata distribucija populacije)
- Više od dvije populacije
  - Jednakost aritmetičkih sredina više od dvije populacije (ANOVA)
  - Jednakost proporcija više od dvije populacije (Hi-kvadrat test)

Ostali testovi (jedna populacija)

- Test hipoteze o pretpostavljenoj vrijednosti varijance normalno distribuirane populacije: Hi-kvadrat test
- Testna veličina empirijski Hi-kvadrat
- $\chi^2 = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2}$
- Uz pretpostavku da je nulta hipoteza istinita, testna veličina pripada Hi-kvadrat distribuciji s $df = n - 1$ stupnjeva slobode
- Tablica 7.6. Bahovec i Erjavec (2015) Statistika; str. 311

Odnos parametara dviju populacija

1. Procjena razlike vrijednosti parametara dviju populacija (jednim brojem i intervalom).
2. Testiranje pretpostavke o razlikama vrijednosti parametara dviju populacija.

- Da bi se proveli postupci procjenjivanja i testiranja hipoteza o razlikama parametara dviju populacija, potrebno je poznavati oblik pripadne sampling-distribucije.

Razlika aritmetičkih sredina dviju populacija: nezavisni uzorci

- Procjene i testiranje provode se pomoću jednostavnih slučajnih uzoraka.
- Pritom je veoma važno jesu li korišteni uzorci podataka **zavisni** ili **nezavisni** i jesu li populacije, iz kojih su uzorci izabrani, normalno distribuirane
- Uzorci su **nezavisni** ako je uzorak opažanja ili mjerenja vezanog za elemente jednog osnovnog skupa neovisan o uzorku opažanja ili mjerenja izabranog iz drugog osnovnog skupa
- Uzorci su zavisni ako su podatci prikupljeni prije i nakon primjene određenog tretmana za iste jedinice uzoraka izabranih iz istog osnovnog skupa, ili se radi o podatcima za jedinice iz skupina sličnih ili povezanih pojedinaca

Test hipoteze o pretpostavljenoj razlici aritmetičkih sredina nezavisnim uzorcima

- Testovi razlike između aritmetičkih sredina dviju normalno distribuiranih populacija provode se pomoću:
  - nezavisnih,
  - zavisnih uzoraka.
- Varijance populacija mogu biti:
  - poznate ili
  - nepoznate.
- Nepoznate varijance mogu biti:
  - jednake ili
  - nejednake.
- Procjenitelj razlike aritmetičkih sredina dviju proizvoljno distribuiranih populacija, ako se koriste veliki uzorci, ima približno normalnu distribuciju.

Testovi hipoteza o pretpostavljenoj razlici sredina dviju normalno distribuiranih populacija kada su varijance poznate

- Izabrana su dva nezavisna slučajna uzorka veličine n_1 i n_2 iz normalno distribuiranih populacija s aritmetičkim sredinama μ_1 , μ_2 i s poznatim varijancama.
- Sampling-distribucija razlike sredina je normalna
- odluka o ishodu testova se donosi s pomoću testne veličine, (empirijskog z –omjera)

Razlika proporcija dviju populacija: veliki nezavisni uzorci

- Moguće je s pomoću jednostavnih slučajnih uzoraka razliku proporcija dviju populacija procijeniti:
  - jednim brojem i
  - intervalom,
- Moguće je i testirati hipotezu da je ta razlika:
  - jednaka,
  - manja ili
  - veća od neke vrijednosti
- Uzorci mogu biti:
  - zavisni i
  - nezavisni.

Usporedba varijanci dviju normalno distribuiranih populacija

- Hipoteze su postavljene u obliku omjera varijanci (ne kao razlike kao što je slučaj s usporedbom aritmetičkih sredina i proporcija)

Dvosmjerni test	Jednosmjerni test na gornju granicu	Jednosmjerni test na donju granicu
$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1, \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	$H_0: \frac{\sigma_1^2}{\sigma_2^2} \leqslant 1, \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$	$H_0: \frac{\sigma_1^2}{\sigma_2^2} \geqslant 1, \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1$

- Testna veličina
- $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$
- je slučajna varijabla koja ima F-distribuciju s ⟦df⟧ _1=n_1-1 stupnjeva slobode u brojniku i ⟦df⟧_2=n_2-1 stupnjeva slobode u nazivniku

Neparametarski testovi

- Ne sadrže pretpostavke o obliku distribucije ili o metričkim svojstvima analiziranih podataka (ili je broj pretpostavki manji/blaži u odnosu na parametarske testove)
- Omogućuju analizu asimetričnih distribucija
- Omogućuju analizu pdataka mjerenih na nominalnoj ili ordinalnoj mjernoj skali

Neparametarski testovi

- **Test predznaka** (engl. *sign test*) za testiranje hipoteze o pretpostavljenoj vrijednosti medijana
- **Wilcoxonov test predznaka** (Uspoređuju se rangovi varijabli)
- **Mann-Whitney U-Test** (za usporedbu medijana dvije populacije)
- **Hi-kvadrat test**

Hi-kvadrat test

- Jedan od najčešće korištenih neparametarskih testova

1. hi-kvadrat test o obliku distribucije populacije (engl. goodness of fit test)
2. hi-kvadrat test o nezavisnosti dviju varijabli (engl. test for independence)
3. Test hipoteze o jednakosti proporcija tri ili više nezavisnih populacija
   - specijalni slučaj hi-kvadrat testa o nezavisnosti dviju varijabli, ako je jedna od varijabli dihotomna

Hi-kvadrat test

- Navedenim testovima želi se utvrditi jesu li odstupanja empirijskih frekvencija od očekivanih frekvencija statistički značajna
- **Empirijske frekvencije** su opažene frekvencije distribucije iz uzorka (engl. observed frequencies).
- **Očekivane (teorijske) frekvencije** (engl. expected frequencies) koje se očekuju pod pretpostavkom da je nulta hipoteza istinita.
- Provođenje hi-kvadrat testa zahtijeva da su vrijednosti promatrane varijable (kvalitativne ili kvantitativne) grupirane u tablicu s modalitetima ili razredima (intervalima), te njima pridruženim apsolutnim frekvencijama.

Hi-kvadrat test

- Hi-kvadrat test može se primijeniti za velike uzorke, ali uz ograničenje da očekivane frekvencije nisu suviše male.
- Ako su očekivane frekvencije suviše male, tada rezultati testa nisu pouzdani (testna veličina bit će neopravdano velika).
- Opće je prihvaćeno pravilo da sve očekivane frekvencije moraju biti jednake ili veće od 5.
- U suprotnom je potrebno združiti frekvencije susjednih modaliteta ili razreda (intervala) ako to ima smisla
- Fisherov test (Fisher's exact test)

Hi-kvadrat test

- Na temelju hi-kvadrat testne veličine mogu se izračunati različiti **koeficijenti "asocijacije"**, kojima se mjeri stupanj (intenzitet) povezanosti među varijablama X i Y .
- koeficijent asocijacije φ ,
- koeficijent kontingence C i
- Cramerov koeficijent V .

Analiza varijance (ANOVA)

- (engl. **An**alysis **o**f **Va**riance – ANOVA)
- Skup statističkih metoda kojima se analizira utjecaj jedne ili više kvalitativnih (kategorijalnih) varijabli na varijacije kontinuirane numeričke varijable.
- Raščlanjivanje varijance promatrane numeričke varijable na komponente koje potječu od različitih izvora varijacija, osnova je za donošenje zaključaka o statističkoj značajnosti promatranih utjecaja.
- ANOVA se primjenjuje u različitim područjima.
- Pomoću analize varijance testira se hipoteza o **jednakosti aritmetičkih sredina *k* populacija** $(k > 2)$.

Analiza varijance (ANOVA)

- Zavisna varijabla Y je predočena kao linearna funkcija jedne ili više kategorijalnih varijabli koje se nazivaju ***faktorima*** ili ***klasifikacijskim varijablama*** (engl. *factors*, *classification variables*).
- Na varijablu Y djeluje jedan ili više faktora, te se ispituje imaju li efekti njihovih razina (tretmana) statistički značajan utjecaj na prosječnu vrijednost varijable Y
- Cilj metode ANOVA je testirati razliku sredina populacija analiziranjem njihovih varijanci

Analiza varijance (ANOVA)

- Analizom varijance može se testirati utjecaj:
  - jednog faktora na zavisnu varijablu (***jednofaktorska analiza varijance***) ili
  - utjecaj dvaju ili više faktora i njihovih interakcija (***višefaktorska analiza varijance***).

Jednofaktorska analiza varijance

- Ispituje se utječe li na varijabilnost promatrane numeričke kontinuirane varijable Y jedan faktor (jedna kategorijalna varijabla).
- Varijabla Y čije se vrijednosti mjere zove se zavisna varijabla (engl. dependent variable, response, output), a faktor koji utječe na varijacije od Y je nezavisna varijabla koja poprima k modaliteta (kategorija), razina ili tretmana (engl. factor levels, treatments).
- Skup vrijednosti varijable Y moguće je razdvojiti na k disjunktnih grupa ili k -populacija.

Jednofaktorska analiza varijance

- $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij},\ i = 1,2,\dots,n_j,\ j = 1,2,\dots,k.$
- Y_{ij}: zavisna varijabla Y za i -to opažanje uz j -tu razinu faktora (j -ti tretman, j -tu populaciju ili j -tu grupu)
- μ: očekivana vrijednost (sredina) zavisne varijable u cijeloj populaciji
- α_j: efekt j -te razine faktora (razlika između očekivane vrijednosti zavisne varijable uz j –ti tretman i očekivane vrijednosti varijable u cijeloj populaciji)
- ε_{ij}: slučajna varijabla kojom se opisuje razlika između opažene i -te vrijednosti j -te grupe i sredine j -te grupe, tj. očekivane vrijednosti zavisne varijable uz j -ti tretman

Pretpostavke ANOVE

1. Iz populacija su izabrani nezavisni jednostavni slučajni uzorci
2. Populacije su normalno distribuirane
3. Varijance svih grupa su jednake (pretpostavka o homogenosti varijance zavisne varijable Y).

Hipoteze

- $H_0: \mu_1 = \mu_2 = \ldots = \mu_k = \mu$
- $H_1: \exists \mu_j \neq \mu, j = 1, 2, \ldots, k.$
- Iz svake populacije se bira jednostavni slučajni uzorak veličine n_j , $j = 1, 2, \ldots, k$
- Provodi se *F*-test

Ishod testa

- 2 moguća ishoda:
- H_0: aritmetičke sredine su jednake
- H_1: aritmetičke sredine nisu jednake.
  - U tom slučaju zanima nas kod kojih populacija se one najviše razlikuju.
  - U tu svrhu provodi se ***Bonferronijev test***.
- Bonferronijev test: Testiraju se razlike sredina svih tretmana (grupa) t –testom na bazi nezavisnih uzoraka

Pretpostavke ANOVE

1. Iz populacija su izabrani nezavisni jednostavni slučajni uzorci
2. Populacije su normalno distribuirane
3. Varijance svih grupa su jednake (pretpostavka o homogenosti varijance zavisne varijable Y).
    - Levenov test homogenosti varijanci

Levenov test

- U nultoj hipotezi testa pretpostavlja se homogenost (jednakost) varijanci.
- Analiza varijance se provodi na osnovi apsolutnih odstupanja opaženih vrijednosti od sredina odgovarajućih grupa.
- Ako je testna veličina (*F*-omjer) statistički značajna uz određenu razinu značajnosti, nulta hipoteza o jednakosti varijanci se odbacuje

Kruskal-Wallisov test

- Ukoliko Levenov test ukazuje na nehomogenost varijanci provodi se Kruskal-Wallisov test.
- Općenito, testovi koji se zasnivaju na pretpostavkama o distribuciji vjerojatnosti u populaciji zovu se parametarskim testovima.
- Ako pretpostavke o distribuciji vjerojatnosti nisu poznate, mogu se koristiti neparametarski testovi, odnosno testovi koji ne ovise o obliku distribucija iz kojih se biraju uzorci.
- Parametarski su testovi superiorniji u odnosu na neparametarske testove, jer imaju veću snagu testa

Kruskal-Wallisov test

- Kruskal-Wallis test: neparametarski test
- testira jednakost medijana u svim populacijama.
- Test se bazira na vrijednostima zavisne varijable transformiranim u rangove, pa se ubraja u testove ranga (engl. rank test).

Kruskal-Wallisov test (pretpostavke)

- 1. Zavisna varijabla je kontinuirana ili redoslijedna.
- 2. Populacije iz kojih se biraju uzorci su jednako distribuirane, osim što se mogu razlikovati u položaju medijana.
- 3. Uzorci su nezavisni (uz napomenu da veličina svakog uzorka mora biti barem 5).
- 4. Uzorci su nezavisni i izabrani su iz jednako distribuiranih populacija.

Kruskal-Wallisov test (hipoteze)

- Nultom se hipotezom pretpostavlja da su medijani svih *k* populacija međusobno jednaki
- Testna veličina