Теорија узорака

21. MAJ ’19.

ΛГлаваш

Проблеми код оцењивања дисперзије оцене

Постоји неколико приступа оцењивању дисперзија оцена непознатих популацијских вредности обележја од интереса. Две основне технике које се, при томе, примењују су:

- техника заснована на аналитичком приступу – „обликовање“ израза за дисперзије оцена
- техника заснована на методима случајних група и методима тзв. репликације, односно поновног узорковања

Аналитички приступ

Код аналитичког приступа оцењивању дисперзија сусрећу се следеће тешкоће:

- сложена рачуница за одређивање вероватноћа укључења другог реда (погледати нпр. Horvitz-Thompson-ову оцену, презентација за 3. час, слајд 16) – то је проблем који је присутан код већине планова (вероватносног) узорковања без понављања
- када је реч о оцењивању популацијске средње вредности, односно тотала, у опсежним истраживањима са више нивоа стратификације, односно поделе на групе, дисперзије оцена ових непознатих параметара израчунавају се за сваки појединачан ниво и акумулирају (погледати нпр. оцене за двоетапни узорак, презентација за 9. час); постратификација односно поступање са неодговором, такође, утичу на дисперзију
- постоји потреба за оцењивањем и других параметара популације (поред средње вредности и тотала) – оцене које се тада користе су, у општем случају, нелинеарне (могу укључивати производе, количнике, степене и др) па се не могу добити тачне (већ, евентуално, апроксимативне) формуле за њихове дисперзије (погледати нпр. оцену популацијског количника, презентација за 4. час, слајд 6)

Метод линеаризације оцене

Када је обим узорка „довољно“ велики, постоји врло приступачан начин поступања са проблемом нелинеарности оцена непознатих популацијских вредности.

Ради се о линеаризацији оцене, развијањем у Тејлоров ред, и прихватању, при одређеним условима, дисперзије линеарне апроксимације оцене као приближне сопственој дисперзији те оцене. Након тога, на уобичајен начин, врши се оцењивање дисперзије линеарне апроксимације оцене на основу узорка.

Овај метод омогућава одређивање прецизности нпр. количничке и регресионе оцене средње вредности, односно тотала главног обележја, као и оцена сложенијих параметара популације као што је коефицијент корелације, линеарни регресиони коефицијент и сл. (погледати додатни материјал за 5. час)

Методи случајних група

- Понављање / копирање плана узорковања ('replicating the survey design')

  Претпоставимо да је план узорковања поновљен R пута на независан начин, под чиме подразумевамо да су сваки пут, након одабира узорка применом фиксираног метода и регистровања вредности обележја од интереса, јединице посматрања из тог узорка враћене у популацију, те су доступне при сваком наредном одабиру узорка. На основу тих R „копија“ узорка затим се формира R независних оцењених вредности непознатог параметра популације, а варијабилност која се појављује међу њима може се искористити за оцењивање дисперзије тачкасте оцене тог непознатог параметра и то на следећи начин.

  Уводе се ознаке:

  θ – непознати параметар

  $\hat{\theta}_r$ – оцењена вредност параметра θ израчуната на основу r-тог „поновљеног“ узорка

$$\tilde{\theta} := \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$$

  Ако је статистика чија је реализација $\hat{\theta}_r$ непристрасна оцена параметра θ, онда је и $\tilde{\theta}$ реализација непристрасне оцене параметра θ, а оцена њене дисперзије дата је са:

$$\frac{1}{R}\cdot\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}_r-\tilde{\theta}\right)^2$$

  ☝ приметити да ова формула има облик количника узорачке дисперзије R независних оцена параметра θ и вредности R

Методи случајних група (наставак)

- Подела узорка на случајне групе ('dividing the sample into random groups')

  Претпоставимо да је одабран (прост) случајан узорак без понављања обима n из популације обима N. Комплетан узорак затим се дели на R дисјунктних група обима по $m := \frac{n}{R}$ јединица посматрања, при чему су групе формиране случајним распоређивањем јединица (нпр. прва група је прост случајан подузорак без понављања обима m из почетног, комплетног узорка; друга група је прост случајан подузорак без понављања обима m одабран из скупа преосталих $(n - m)$ јединица из почетног узорка итд). Идеја је да на свакој оваквој (псеудо)случајној групи имамо умањену верзију истраживања, која ће у потпуности одражавати установљен план узорковања. Ипак треба приметити да групе нису међусобно независне „копије“ јер се појединачна јединица посматрања може појавити у тачно једној групи. ☝ ако је $n \ll N$ оне се могу сматрати независним копијама

  Овде се као тачкаста оцена непознатог параметра θ предлаже:

$$\tilde{\theta} := \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r$$

  где је $\hat{\theta}_r$ – оцена параметра θ израчуната на основу r-те групе. Ако је θ нелинеаран параметар, оцена $\tilde{\theta}$, у општем случају, неће бити једнака тачкастој оцени $\hat{\theta}$, која је истог облика као $\hat{\theta}_r$ али је формирана на основу почетног, комплетног узорка. Без обзира на то, (помало прецењена) оцена $D\hat{\theta}$ може се дати формулом:

$$\frac{1}{R}\cdot\frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}\right)^2$$

нпр. код оцењивања популацијског количника:

$\tilde{B} = \frac{1}{R}\sum_{r=1}^{R}\frac{\bar{Y}_r}{\bar{X}_r}$

$\hat{B} = \frac{\bar{Y}}{\bar{X}}$

Методи репликације и поновног узорковања

- Jackknife

Овај метод уопштава метод поделе узорка на случајне групе, на тај начин што је дозвољено преклапање група. У наставку ће бити описан тзв. delete-1 jackknife (назив потиче одатле што се „копије“ група формирају избацивањем по једне јединице посматрања из почетног, комплетног узорка).
Претпоставимо да је одабран (прост) случајан узорак $\mathcal{S}$ без понављања обима n из популације обима N. Број група ће, стога, овде бити једнак n. Нека је $\hat{\theta}$ тачкаста оцена параметра θ формирана на основу почетног, комплетног узорка, а $\hat{\theta}_{(l)}$ оцена истог облика као и $\hat{\theta}$ али у којој не фигурише јединица $l \in \mathcal{S}$.
Оцена $D\hat{\theta}$ може се дати формулом:

$$\frac{n-1}{n}\sum_{l\in\mathcal{S}}\left(\hat{\theta}_{(l)}-\tilde{\theta}\right)^2$$

где је

$$\tilde{\theta}=\frac{1}{n}\sum_{l\in\mathcal{S}}\hat{\theta}_{(l)}$$

→ jackknife оцена пристрасности оцене $\hat{\theta}$ параметра θ дата је са:

$$(n-1)(\tilde{\theta}-\hat{\theta})$$

па се овај метод може користити и за смањење пристрасности оцена

нпр:

$$\hat{\theta}=\frac{1}{n}\sum_{k\in\mathcal{S}}y_k$$

$$\hat{\theta}_{(l)}=\frac{1}{n-1}\sum_{k\in\mathcal{S}\setminus\{l\}}y_k$$

Методи репликације и поновног узорковања (наставак)

- Bootstrap

  Овај метод поновног узорковања популаризован је '80-их година. Премда је изузетно једноставан за имплементацију, његова примена не би била (тако лако) изводљива и ефикасна без напредне рачунарске технологије. Наиме, емпиријски (непараметарски) bootstrap је у основи симулациона процедура која се извршава на рачунару и која се заснива на идеји да се поновним узорковањем ('resampling') јединица из доступног узорка (третирајући га као популацију) генерише велики број других узорака, који ће затим бити искоришћени за оцењивање прецизности различитих статистика.
  Претпоставимо да је одабран (прост) случајан узорак $\mathcal{S}$ са понављањем обима n из популације обима N. Кључни појам који се користи у наставку је bootstrap узорак $\mathcal{S}^*$, који представља прост случајан узорак са понављањем истог обима n одабран из почетног узорка $\mathcal{S}$ (као да је $\mathcal{S}$ популација). ☝ он се може посматрати као узорак из емпиријске функције расподеле узорка $\mathcal{S}$
  Ако је узорак $\mathcal{S}$ заиста репрезентативан за посматрану популацију (тј. ако је његова емпиријска функција расподеле добро апроксимира функцију расподеле обележја од интереса на читавој популацији), онда очекујемо да ће се bootstrap узорци „понашати“ као да су одабрани директно из популације.

Bootstrap (наставак)

Нека је $\hat{\theta} = \hat{\theta}(\mathcal{S})$ тачкаста оцена непознатог параметра θ, а $\hat{\theta}^* := \hat{\theta}(\mathcal{S}^*)$ тзв. bootstrap „копија“ $\hat{\theta}$, која се добија применом исте функције на bootstrap узорак уместо на почетни узорак.

Bootstrap алгоритам за оцењивање дисперзије оцене $\hat{\theta}$ има следеће кораке:

- одабрати B независних bootstrap узорака $\mathcal{S}_1^*, \mathcal{S}_2^*, \ldots, \mathcal{S}_B^*$ идеално: $B \to +\infty$
- одредити реализовану вредност $\hat{\theta}^*$ за сваки bootstrap узорак одабран у првом кораку: $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ $(\hat{\theta}_l^* := \hat{\theta}(\mathcal{S}_l^*))$
- оценити дисперзију оцене $\hat{\theta}$ узорачком дисперзијом B bootstrap „копија“ $\hat{\theta}$, тј. применом формуле:

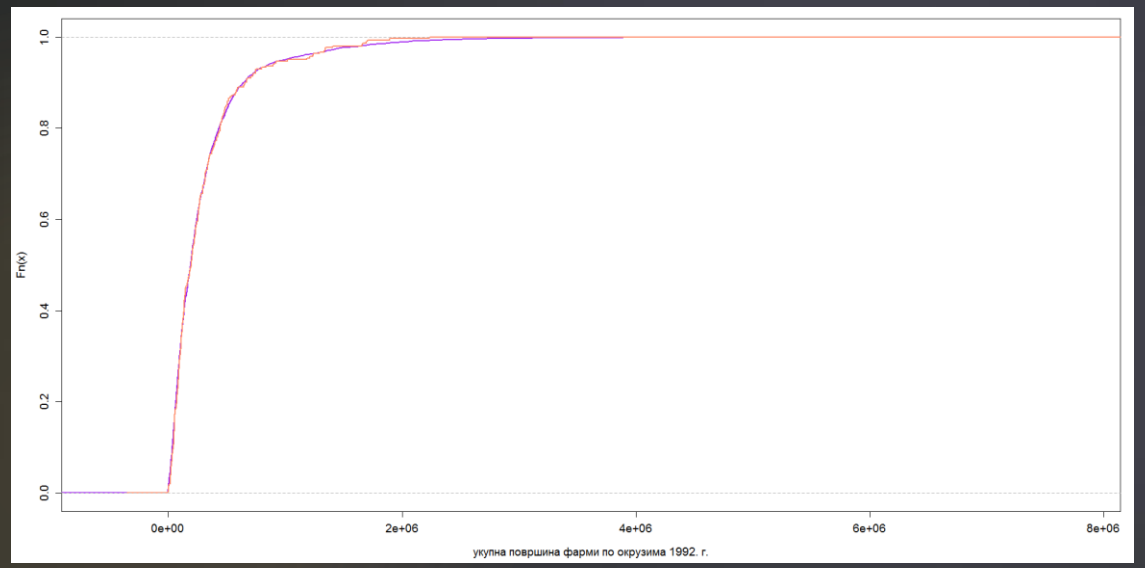
$$\frac{1}{B-1}\sum_{l=1}^{B}\left(\hat{\theta}_l^* - \tilde{\theta}^*\right)^2 \quad \text{где је} \quad \tilde{\theta}^* = \frac{1}{B}\sum_{l=1}^{B}\hat{\theta}_l^*$$

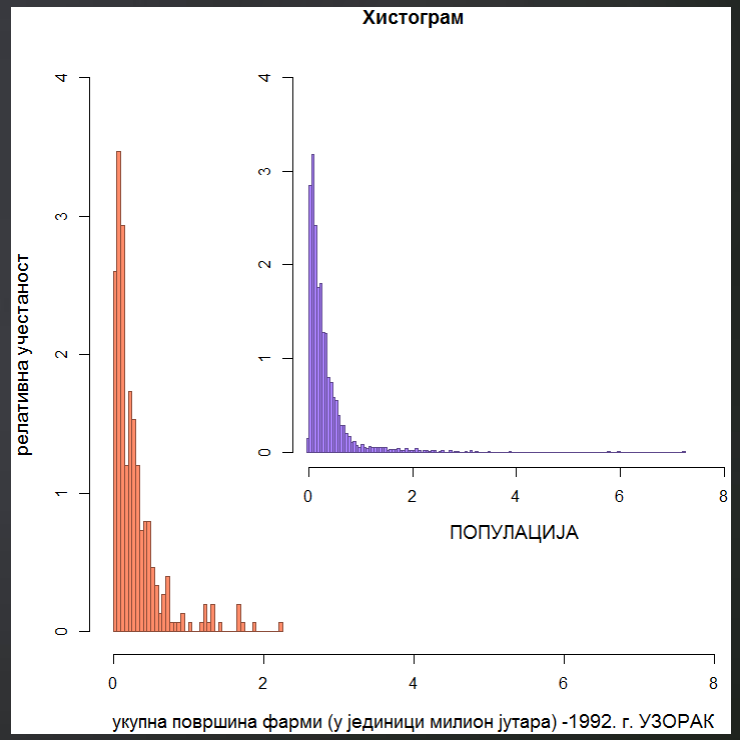
Пример – оцењивање медијане

Погледати пример из презентације за 4. час, слајд 11, са подацима из Пописа пољопривреде у САД за 1992. г.

Главно обележје Y је укупна површина фарми (по окрузима, који представљају јединице узорковања) 1992. г. На основу простог случајног узорка (agsrs) обима $n = 300$ оцењује се популацијска медијана и примењује техника bootstrap оцењивања дисперзије ове оцене.

☝ Почетни узорак из базе agsrs одабран је методом SRSWOR,
али ћемо ми са њим поступати као да је прост случајан узорак са враћањем,
јер је обим узорка релативно мали ($\approx 9.75\%$) у односу на обим популације





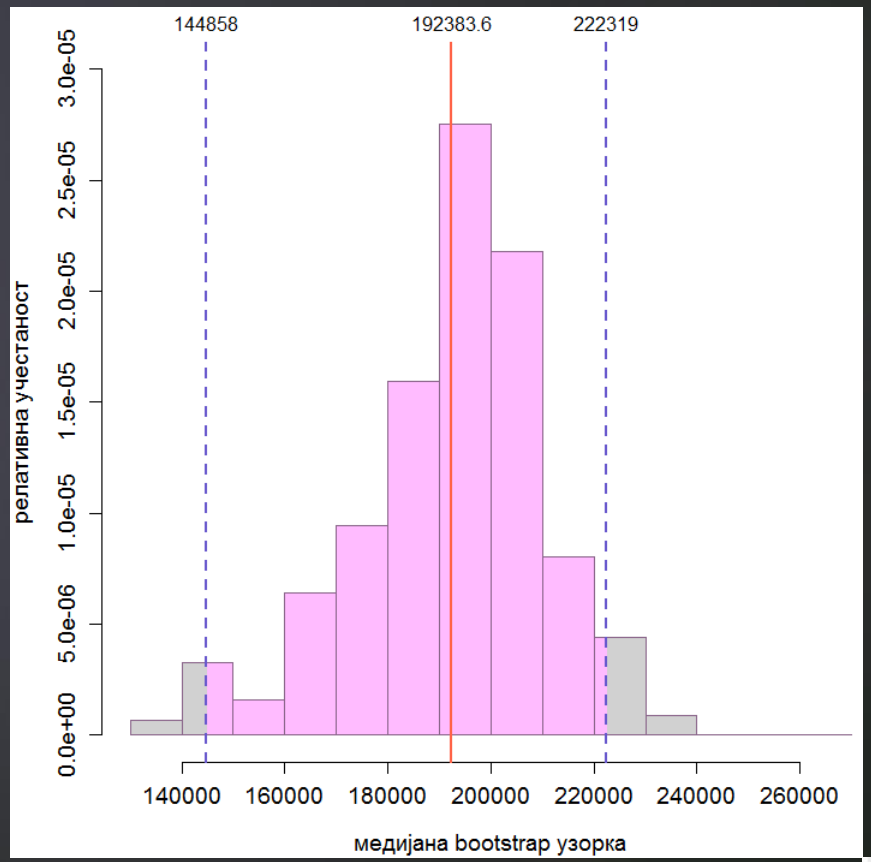
На горњој слици, лево, приказани су графици:
функције расподеле обележја Y на читавој популацији
емпиријске функције расподеле обележја Y на узорку.
Судећи по слици десно очекујемо да bootstrap узорци обима 300 из почетног узорка добро „имитирају“ просте случајне узорке са понављањем који би се бирали из читаве популације.

ЛГлаваш

Пример – оцењивање медијане – наставак

С обзиром на врло изражену (позитивну) асиметрију података из узорка, која се лако уочава са хистограма, медијана представља врло значајан показатељ централне тенденције (и информативнија је од нпр. средње вредности).
Узорачка медијана једнака је: 196717 (јутара)
Bootstrap оцена стандардне грешке узорачке медијане има реализовану вредност (за $B = 3000$): 18678.55

На слици је приказан хистограм bootstrap копија узорачке медијане. Расподелом вероватноћа ових података (тзв. bootstrap расподела $\hat{\theta}^*$) суштински се оцењује узорачка расподела медијане. На основу ње могу се формирати (апроксимативни) bootstrap интервали поверења.
Нпр. за одређивање двостраног 95% интервала поверења за популацијску медијану потребно је одредити 2.5 перцентил и 97.5 перцентил bootstrap расподеле, па је овде такав интервал $[144858, 222319]$.



- Ако је bootstrap расподела $\hat{\theta}^*$ приближно нормална расподела вероватноћа, онда се апроксимативни интервали поверења могу формирати и на тај начин. Овде: $[161477.8, 231956.2]$

ЛГлаваш

Невероватносно узорковање

Често се у истраживањима користе и узорци који нису формирани случајним одабиром јединица популације. Таквом одабиру јединица приступа се из различитих разлога и у разне сврхе.

При одлучивању о избору плана узорковања, важно је знати да узорци који нису одабрани у складу са извесном расподелом вероватноћа, дефинисаном на колекцији свих могућих узорака, нису погодни за уопштавање добијених резултата и закључака на комплетну популацију (што некада није ни циљ истраживања).

Извођење (и евентуално уопштавање) закључака на основу невероватносних узорака изискује примену другачијих процедура него што је то случај код вероватносних планова узорковања. Последњих година, пре свега услед убрзаног технолошког развоја, појавили су се неки нови приступи невероватносном узорковању, који се најопштије могу назвати приступима заснованим на моделу.

☝ "All models are wrong, but some are useful!" George E. P. Box

Пригодни узорак и узорак добровољаца

Пригодни узорак ('convenience sampling') сачињен је од јединица популације које су расположиве и лако доступне истраживачу. Дакле, примарни критеријум за одабир је лакоћа узорковања (једноставност и ефикасност његовог спровођења, која је обично праћена економичношћу). Према томе, постоје и јединице у популацији за које не постоји никаква шанса да буду одабране у узорак. Овако формиран узорак је врло ретко репрезентативан, без обзира на свој обим.
Велика мана овог плана узорковања јесте што су непознати смер и величина разлика између вредности добијених испитивањем узорка и стварних вредности у целој популацији. Такође, не постоји могућност квантификовања грешке резултата истраживања.
Пригодни узорак се може користити у експлоративним истраживањима, где генерализација резултата није у првом плану, јер су то обично почетна истраживања за серију накнадних, у којима би требало, на коректним узорцима, потврдити или одбацити почетне налазе.
Примери: анкетирање посетилаца тржних центара или других јавних места ('mall intercept interviewing'); анкетирање запослених из узорка пословних објеката који се налазе у непосредној близини седишта организације која је задужена за прикупљање података у истраживању и свако слично несистематско регрутовање појединаца; тражење да појединац који је посетио извесну web страницу попуни анкету; слање кратког упитника уз рачун за одређену услугу итд).

Иако је код сваког истраживања кључна добра воља узорковане јединице да се одазове истом, под узорком добровољаца ('volunteer sampling') подразумева се пре свега узорак сачињен од појединаца који су се сами пријавили (‘self-selection') за учешће у истраживању. Код њега може бити изузетно изражена пристрасност испитаника.

Код оба типа узорка често остаје нејасно које све јединице чине популацију.

Намерни узорак и узорак „снежне грудве“

Намерни узорак ('purposive sampling') сачињен је од јединица популације одабраних у узорак на основу својстава која су прецизно дефинисана проблемом, предметом и циљевима истраживања (у људској популацији нпр. на основу одређених способности, активности, жеље појединаца да учествују у истраживању, знања из области од интереса итд), при чему истраживач тежи да постигне „репрезентативност“ узорка.
Узорковање се, ипак, заснива на субјективном просуђивању истраживача, често уз примену стручних знања из области која се проучава. Та стручна знања о испољавању појаве која се испитује на датој популацији требало би прво да укажу на то која својства јединица популације је важно да буду заступљена у узорку (у том смислу би требало и схватити репрезентативност). Затим се, на основу тога, бира узорак који ће представљати „попречни пресек“ популације.
Намерни узорак је погодан за експлоративна истраживања, посебно за третирање екстремних случајева у популацији. Прецизнији је у односу на пригодни узорак.

Узорак „снежне грудве“ ('snowball sampling') користи се за аналитичка истраживања, и то искључиво у људској популацији. Формира се тако што се одабере известан број испитаника, који током прикупљања података указују на нове испитанике, који би могли бити укључени у узорак. Ти нови испитаници даље указују на наредне итд, па је сам поступак долажења до нових испитаника сличан ефекту снежне грудве, одакле потиче и назив овог типа невероватносног узорковања.
Узорак „снежне грудве“ прикладан је за испитивање својстава која се у популацији ретко јављају, тј. за идентификовање појединаца који су иначе „тешко уочљиви“.

Квотни узорак

Квотни узорак ('quota sampling'), на известан начин, представља лошију верзију стратификованог узорка. Након дефинисања циљне популације потребно је да истраживач поседује (додатну, помоћну) информацију за сваку јединицу у њој (о неком варијабилном својству – често и о више таквих својстава, које је интересантно / значајно за појаву која се испитује, и које је, стога, одређено проблемом, предметом и циљевима истраживања), на основу чега ће популација бити раслојена на подскупове јединица (тзв. потпопулације). Потом се одређује обим сваке потпопулације, обим узорка и квоте, тј. број јединица сваке потпопулације које треба испитати, али тако да распоред узорка буде пропорционалан (броју јединица у потпопулацијама у односу на обим читаве популације). Поштовањем квота тежи се постизању репрезентативности узорка. Коначно, одабир потенцијалних јединица из сваке потпопулације у узорак врши се слободним просуђивањем и одлучивањем истраживача (што може нарушити репрезентативност).
Овакво узорковање није ресурсно захтевно (у погледу утрошка времена, новца, ангажовања и др), па се, због свега тога, сматра најзначајнијим планом невероватносног узорковања.