Стратификовани случајан узорак

5. и 6. час

1. април 2016.

Стратификовани случајан узорак

Стратификован узорак примењује се, пре свега, онда када је потребно повећати прецизност оцена параметара, односно смањити грешке узорка. Други разлози могу бити: економичност, једноставност истраживања и слично. У неким ситуацијама истраживање је могуће спроводити једино по стратумима.

Стратификација (раслојавање) је подела популације на потпопулације, стратуме, при чему би требало формирати релативно хомогене, међу собом разграничене стратуме, што значи да вредности обележја, које је предмет истраживања, буду приближне на јединицама у сваком стратуму, а да се вредности обележја јединица из различитих стратума међусобно битно разликују.

Читава популација се класификује у стратуме на основу неких додатних, претходно прикупљених информација. Као критеријум за стратификацију користи се нека (једна или више њих) карактеристика популације, за коју се сматра да је са посматраним обележјем у корелацији.

Заправо, повећање прецизности оцене зависи од хомогености јединица у оквиру стратума и на њега, у великој мери, утиче начин стратификације. Зато се, природно, постављају питања:

- како формирати стратуме
- како одредити број стратума
- како распоредити узорак по појединим стратумима

Након извршене стратификације узорци, унапред одређеног обима, бирају се унутар сваког стратума. При томе, узорци се бирају међусобно независно из различитих стратума и није неопходно користити исти план узорковања за све стратуме.

Примери ситуација, код којих би било погодно користити стратификацију:

- **јединице популације**: пољопривредна газдинства
  обележје: принос пшенице
  стратификација: укупна површина обрадивог земљишта по фарми
- **јединице популације**: области у географским регионима
  обележје: број домаћинстава
  стратификација: густина насељености; рељеф
- **јединице популације**: људи
  обележје: разна
  стратификација: пол, старост, образовање, верска припадност, етничка припадност, област живљења, социјално-економски фактори и сл.

Уопште, стратификацију је погодно користити онда када је приметно да обележје од интереса има различите средње вредности у различитим подгрупама јединица популације.

Захтева се:

- стратуми морају бити међусобно дисјунктни, тј. свака јединица популације мора припадати тачно једном стратуму
- стратуми морају "покривати" целу популацију, тј. не сме се појавити јединица која није укључена ни у један стратум
- требало би да стратуми буду интерно хомогени, а да се међусобно значајно разликују
- број стратума може бити већи или мањи, с тим што се због мерења прецизности оцене, захтева да број јединица у сваком стратуму не сме бити мањи од две јединице

Такође, инсистира се на томе да начин поделе популације на стратуме буде, што је више могуће, "природан".

Предности:

- могућност да се не само оцене параметри на целој популацији, него и да се донесу закључци на нивоу, тј. унутар самих стратума, и да се изврши поређење по стратумима
- могућност да истраживач сам контролише величине узорка унутар сваког стратума
- повећање прецизности оцене у смислу смањења дисперзије оцена (нпр. у односу на узорак SRSWOR истог обима)
- повећање репрезентативности узорка, јер омогућава да елементи сваког стратума буду заступљени у финалном узорку
- могућност да истраживач користи различите планове узорковања на различитима стратумима, у зависности од сопствених потреба и доступности информација
- јефтиније је

Ознаке:

- L - број стратума у популацији
- N_h - број јединица у h-том стратуму, $h = 1, 2, ..., L$
- Y_{hj} - вредност обележја y j-те јединице h-тог стратума, $h = 1, 2, ..., L$, $j = 1, 2, ..., N_h$
- n_h - величина узорка који се бира из h-тог стратума
- Y_h - тотал обележја h-тог стратума
- $\bar{Y}_h$ - средина обележја h-тог стратума
- y_{hj} - вредност обележја y j-те јединице одабране у узорак у h-том стратуму
- $\bar{y}_h$ - узорачка средина обележја за h-ти стратум

Важи:

$$N = \sum_{h=1}^{L} N_h, \quad n = \sum_{h=1}^{L} n_h$$

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} [Y_{hj} - \bar{Y}_h]^2, \quad s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} [y_{hj} - \bar{y}_h]^2$$

Ако је $\hat{Y}_h$, $h = 1, 2, ..., L$, непристрасна оцена тотала обележја Y_h h-тог стратума, тада је непристрасна оцена тотала обележја популације

$$\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h,$$

а оцена њене дисперзије је

$$v[\hat{Y}_{st}] = \sum_{h=1}^{L} v[\hat{Y}_h],$$

где су $v[\hat{Y}_h]$ непристрасне оцене дисперзија $V[\hat{Y}_h]$.

Ако је $\hat{\bar{Y}}_h$, $h = 1, 2, ..., L$, непристрасна оцена средине обележја $\bar{Y}_h$ h-тог стратума, тада је непристрасна оцена средине обележја популације

$$\hat{\bar{Y}}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h \hat{\bar{Y}}_h.$$

Ако се из сваког стратума бира прост случајан узорак, онда је непристрасна оцена тотала обележја популације је

$$\hat{Y}_{st} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hj},$$

а непристрасна оцена њене дисперзије је

$$v[\hat{Y}_{st}] = \sum_{h=1}^{L} \frac{N_h^2}{n_h} s_h^2,$$

односно

$$v[\hat{Y}_{st}] = \sum_{h=1}^{L} \frac{N_h^2(N_h - n_h)}{N_h n_h} s_h^2.$$

Фракција узорка у h-том стратуму је $f_h = \frac{n_h}{N_h}$.

Одређивање обима узорка по стратумима

Када је већ одређен и фиксиран обим узорка, треба приступити одлучивању о обиму узорка n_h за сваки стратум појединачно, $h = 1, 2, ..., L$.

У пракси се за решавање овог проблема обично користи нека од две популарне технике:

1. пропорционални распоред
2. оптималан распоред

Одређени распоред обима узорка по стратумима примењује се, пре свега, у циљу смањења дисперзије. Међутим, и други чиниоци условљавају размештај обима узорка.

Пропорционални распоред за фиксну величину обима узорка

Код пропорционалног распореда, број јединица које се бирају у узорак из појединог стратума, пропорционалан је броју јединица у том стратуму, тј. $n_h = \frac{n}{N} N_h$, $h = 1, 2, ..., L$.

Код пропорционалног распореда и стратификованог СУ, непристрасна оцена тотала обележја популације дата је са

$$\hat{Y}_{st} = \frac{N}{n} \sum_{h=1}^{L} \sum_{j=1}^{n_h} y_{hj}.$$

Ова техника даје обиме узорака по стратумима онда када је унапред познат обим целог узорка и не узима у обзир трошкове. Међутим, трошкови су увек значајно ограничење при организовању било каквог истраживања. Зато је од интереса размотрити пропорционални распоред за задати укупан трошак.

Пропорционални распоред са фиксним трошковима истраживања

Нека је c_h, $h = 1, 2, ..., L$, трошак прикупљања информације од једне јединице из h-тог стратума. Ови трошкови могу се значајно разликовати међу стратумима.

Укупан трошак истраживања је

$$C = C_0 + \sum_{h=1}^{L} c_h n_h,$$

где је C_0 општи (сталан) трошак.

Пропорционални распоред за дати трошак дат је са

$$n_h = \frac{C - C_0}{\sum_{h=1}^{L} c_h N_h} N_h,$$

а укупан обим узорка је, тада, једнак

$$n = \frac{C - C_0}{\sum_{h=1}^{L} c_h N_h} N.$$

Непропорционални распореди

Претходно описана техника пропорционалног распореда не узима у разматрање ниједан други аспект предмета истраживања, осим величине стратума (тј. броја јединица у стратуму). Она у потпуности игнорише унутрашњу структуру стратума.

Зато су предложене и шеме распореда, које воде рачуна о поменутом.

У пракси се користе две шеме распореда које минимизирају дисперзију оцена. Како је минимална дисперзија оптимално својство оцене, овакви распореди се називају оптималним.

- Neyman-ов распоред
- Cost Optimum Allocation

Neyman-ов распоред

Neyman-ов распоред минимизира дисперзију оцене за познат и фиксиран обим целог узорка.

Код стратификованог случајног узорка без понављања, дисперзија оцене тотала обележја $\hat{Y}_{st}$ износи

$$V[\hat{Y}_{st}] = \frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h\right)^2 - \sum_{h=1}^{L} N_h S_h^2.$$

Циљ је одредити $n_1, n_2, ..., n_L$, који минимизирају наведену дисперзију, под условом да важи $\sum_{h=1}^{L} n_h = n$. Рачуницом се добија

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} n.$$

Cost Optimum Allocation

Овај распоред минимизира дисперзију оцене за познат и фиксиран укупан трошак истраживања. Рачуницом се добија

$$n_h = \frac{\frac{N_h S_h}{\sqrt{c_h}}(C - C_0)}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}},$$

а укупан обим узорка је једнак

$$n = \frac{(C - C_0)\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{c_h}}}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}.$$

Што је већа варијабилност посматраног обележја у одређеном стратуму или што је већи обим стратума или што је јефтиније узорковање унутар стратума, требало би узети већи узорак из стратума.

Одређивање обима узорка за задату тачност

Нека је d апсолутна грешка оцене непознатог параметра и α праг значајности. Код стратификованог случајног узорка без понављања потребан обим узорка за оцену тотала обележја популације је:

- код равномерног распореда $(n_h = \frac{n}{N})$

$$n = \frac{n_0}{1 + \frac{z^2}{d^2}\sum_{h=1}^{L} N_h S_h^2}, \quad n_0 = \frac{Lz^2}{d^2}\sum_{h=1}^{L} N_h^2 S_h^2$$

- код пропорционалног распореда $(n_h = \frac{nN_h}{N})$

$$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}}, \quad n_0 = \frac{Nz^2}{d^2}\sum_{h=1}^{L} N_h S_h^2$$

- код Neyman-овог распореда $(n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h^2} n)$

$$n = \frac{n_0}{1 + \frac{z^2}{d^2}\sum_{h=1}^{L} N_h S_h^2}, \quad n_0 = \frac{z^2}{d^2}\left(\sum_{h=1}^{L} N_h S_h\right)^2$$

Главни разлог због кога би предност требало дати стратификованом узорковању јесте што овај план узорковања даје прецизније оцене него (прост) СУ.

Генерално, план узорковања може се вредновати поређењем дисперзије одговарајуће оцене непознатог параметра, добијене тим планом узорковања, са дисперзијом исте оцене добијене код СУ без понављања (ради се са узорком истог обима). Количник те две дисперзије је ефекат дизајна (design effect) – DEFF.

$$DEFF[\hat{Y}_{st}] = \frac{V[\hat{Y}_{st}]}{V[\hat{Y}_{srs}]},$$

односно

$$DEFF[\hat{\bar{Y}}_{st}] = \frac{V[\hat{\bar{Y}}_{st}]}{V[\hat{\bar{Y}}_{srs}]}$$

Ако DEFF има вредност мању од 1 то указује да је стратификован случајни узорак ефикаснији од СУ без понављања истог обима. Ако има вредност већу од 1, ефикаснији је СУ без понављања истог обима.

Поређење квалитета оцена код стратификованог и простог случајног узорка

Претпоставимо да је обим узорка n исти, унапред одређен и фиксиран.

Такође, претпоставимо да је се фактор корекције популације и фактор корекције за стратуме могу занемарити, па упоређујемо дисперзије:

- V_{srs} дисперзија за прост случајан узорак
- V_{prop} дисперзија за стратификовани узорак са пропорционалним распоредом
- V_{opt} дисперзија за стратификовани узорак са оптималним распоредом

Тада важи:

$$V_{opt} \leq V_{prop} \leq V_{srs}.$$

Избор и формирање стратума

Један од основних проблема, који се јављају код стратификованог узорка, тиче се питања броја стратума, а посредно и величине стратума, тј. броја јединица унутар стратума.

Мали број стратума може довести до значајне варијабилности, тј. одступања у вредностима обележја јединица унутар истог стратума.

Велики број стратума отежава рад и знатно повећава трошкове истраживања.

Јасно је да би стратуме требало формирати тако да имају што већу "хомогеност", тј. тако да је S_h^2 у сваком стратуму, што мање, $h = 1, 2, ..., L$. Самим тим и мали обим узорка по стратуму обезбеђује довољну прецизност оцена.

Избор и формирање стратума

Најбоље би било да се стратификација врши директно на основу вредности самог обележја које се испитује.

Међутим, стратификација (раслојавање) по вредностима обележја које се изучава је ретко изводљива, или је чак и бесмислена, јер захтева познавање свих вредности обележја популације.

Ипак, стратификација по самом обележју ("одокативно") је понекад прилично једноставна.

Стратификација се најчешће врши према неком обележју за које постоји основана индиција да је у корелацији са испитиваним обележјем. При томе, очекује се да хомогеност у стратумима у односу на "помоћно" обележје значи и хомогеност вредности посматраног обележја.

Оцена популацијске пропорције

Претпоставимо да је обележје од интереса y, заправо индикатор функција која указује на то да ли одговарајућа јединица популације припада одређеном нивоу посматране категоричке променљиве или не.

Код стратификованог СУ без понављања, непристрасна оцена популацијске пропорције је

$$\hat{p}_{st} = \frac{1}{N} \sum_{h=1}^{L} N_h p_h,$$

где је $p_h = \bar{y}_h$ релативна учестаност припадања датом нивоу у h-том стратуму.

Дисперзија ове оцене може се оценити са

$$v[\hat{p}_{st}] = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{p_h(1-p_h)}{n_h - 1}(1 - f_h).$$

Пример

`API` популација (у пакету `survey`) садржи податке о индексу академског успеха у Калифорнији, који је рачунат на основу тестова, које су решавали ученици калифорнијских школа. Поред података о академским постигнућима ученика по школама, на располагању су и вредности различитих друштвено-економских обележја.

Ови подаци се интензивно користе за илустрацију рада софтвера намењеног анализи података при истраживањима (Academic Computing Services at the University of California, Los Angeles).

```
> install.packages("survey")
> library(survey)
> data("api")
> apipop
```

Пример

Стратификован случајан узорак без понављања обима 200 школа, из ове популације, смештен је у базу података `apistrat`.

Стратификација је вршена на основу нивоа школовања (тј. на основу вредности обележја `stype`), где је

- $n_E = 100$ `elementary schools`
- $n_M = 50$ `middle schools`
- $n_H = 50$ `high schools`

Распоред узорка је направљен на основу следеће идеје:
како су у Калифорнији `high schools` обично веће од `middle schools` односно `elementary schools`, ако би се десило да СУ без понављања садржи више `high schools` то би водило ка "прецењеној" средини и тоталу броја уписаних ученика, док ако би се десило да случајан узорак садржи мање `high schools` то би водило ка "потцењеној" средини и тоталу броја уписаних ученика.

Пример

Следећим кодом "описује" се овај план истраживања R-у.

- Аргументом `strata= stype` назначена је променљива по којој је извршена стратификација.
- Променљива `fpc` у овој бази података чува обиме стратума, а не обим читаве популације ($N_E = 4421, N_M = 1018, N_H = 755$).
- Променљива `pw` у бази садржи "тежине" узорковања (sampling weights).

```
> dstrat <- svydesign(id=~1,strata=~stype,weights=~pw,data=apistrat,fpc=~fpc)
> dstrat
Stratified Independent Sampling design
svydesign(id = ~1, strata = ~stype, weights = ~pw, data = apistrat,
    fpc = ~fpc)
```

Пример

```
> summary(dstrat)

Stratified Independent Sampling design
svydesign(id = ~1, strata = ~stype, weights = ~pw, data = apistrat,
    fpc = ~fpc)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02262 0.02262 0.03587 0.04014 0.05339 0.06623
Stratum Sizes:
             E  H  M
obs        100 50 50
design.PSU 100 50 50
actual.PSU 100 50 50
Population stratum sizes (PSUs):
   E    H    M
4421  755 1018
Data variables:
 [1] "cds"      "stype"    "name"     "sname"    "snum"     "dname"    "dnum"     "cname"
 [9] "cnum"     "flag"     "pcttest"  "api00"    "api99"    "target"   "growth"   "sch.wide"
[17] "comp.imp" "both"     "awards"   "meals"    "ell"      "yr.rnd"   "mobility" "acs.k3"
[25] "acs.46"   "acs.core" "pct.resp" "not.hsg"  "hsg"      "some.col" "col.grad" "grad.sch"
[33] "avg.ed"   "full"     "emer"     "enroll"   "api.stu"  "pw"       "fpc"
```

Пример

Након креирања жељеног `survey design object`-а могуће је проследити га, уз одговарајућу формулу (у зависности од тога шта је потребно израчунати) функцијама:

- `svymean()`
- `svytotal()`
- `svyratio()`
- `svyvar()`
- `svyquantile()`

```
> svytotal(~enroll, dstrat)
 total SE
enroll 3687178 114642
> (m <- svymean(~enroll, dstrat))
 mean SE
enroll 595.28 18.509
```

Пример

У истом пакету налази се и СУ без понављања изабран из API популације и смештен је у базу података apisrs.

```
> dsrs <- svydesign(id=~1, fpc=~fpc, data=apisrs)
> svytotal(~enroll, dsrs)
 total SE
enroll 3621074 169520
> svymean(~enroll, dsrs, deff=T)
 mean SE DEff
enroll 584.610 27.368 1
```

Функције svymean() и svytotal() могу се примењивати и на факторе. У том случају биће креиране табеле са оцењеним релативним, односно апсолутним фреквенцијама за сваки ниво фактора.

```
> svytotal(~stype, dsrs)
 total SE
stypeE 4397.74 196.00
stypeH 774.25 142.85
stypeM 1022.01 160.33
```

Пример

У оквиру истог позива функције `svymean()` или `svytotal()` може се вршити анализа више променљивих и рачунати разлике између резултата.

Следећим кодом израчуната је оцена средње вредности обележја индекс академског успеха (`Academic Performance Index`) за $1999.$ и $2000.$ годину.

```
> (means <- svymean(~ api00+api99, dsrs))
 mean SE
api00 656.59 9.2497
api99 624.68 9.5003
```

Позивом функције `svycontrast()` рачуна се разлика између ових средина по формули:

$$(1 \times 2000.mean) + (-1 \times 1999.mean).$$

```
> svycontrast(means, c(api00=1, api99=-1))
 contrast SE
 contrast 31.9 2.0905
> #alternativna notacija: svycontrast(means, quote(api00-api99))
```

Sampling weights - "тежине" узорковања

Ако се одабере СУ без понављања од 3500 хиљаде људи из земље Недођије (која нпр. има популацију од 35 милиона људи), онда свака особа има вероватноћу укључења у узорак $\pi_i = 0.0001$.

Дакле, сваки човек који се налази у узорку, репрезентује 10000 својих сународника.

Фундаментална статистичка идеја, у позадини закључивања на основу било ког плана узорковања, јесте да јединица узоркована са вероватноћом укључења репрезентује тачно јединица популације. Вредност назива се "тежина" узорковања – sampling weight.

Пакет `sampling`

- `strata()`
  Служи за избор стратификованог узорка са једнаким, односно неједнаким вероватноћама.
- `inclusionprobastrata()`
  Израчунава вероватноће укључења првог реда код стратификованог плана узорковања. Вероватноће укључења једнаке су за све јединице унутар истог стратума.
- `HTstrata(y,pik,strata,description=FALSE)`
  Израчунава Horvitz-Thompson оцена тотала популације за стратификовани узорак.

Оцењивање унутар потпопулација

Најједноставнији приступ јесте коришћењем функције `svyby()`, којом се могу израчунати жељене оцене за скуп потпопулација.

Идеја се може применити како код стратификованог (случајног) узорка, тако и у ситуацијама када треба анализирати подгрупе јединица популације, које нису стратуми.

- **Анализа по стратумима**

```
> (tot <- svyby(~enroll, by=~stype, design=dstrat, svymean))
 stype enroll se
E E 416.78 16.41740
H H 1320.70 91.70781
M M 832.48 54.52157
```

- **Анализа подгрупа популације које нису стратуми**

  Променљива `emer` у `API` популацији чува проценат наставника који имају сертификат за држање наставе само у хитним ситуацијама (`emergency teaching certification`).

Око 20% школа нема наставнике са овим сертификатом и отприлике исти број школа, има више од 20% наставника са овим сертификатом. Следећим кодом оцењена је средња вредност индекса академског успеха и укупан број ученика у обе подгрупе.

```
> emerg_high <- subset(dstrat, emer>20)
> emerg_low <- subset(dstrat, emer==0)
> svymean(~api00+api99, emerg_high)
 mean SE
api00 558.52 21.708
api99 523.99 21.584
> svymean(~api00+api99, emerg_low)
 mean SE
api00 749.09 17.516
api99 720.07 19.061
> svytotal(~enroll, emerg_high)
 total SE
enroll 762132 128674
> svytotal(~enroll, emerg_low)
 total SE
enroll 461690 75813
```

Функција `subset()` искоришћена је како би се креирао survey design object који представља потпопулацију.

Постстратификација (Poststratification)

Било је говора о повећању прецизности оцена непознатих параметара, коришћењем додатних података о популацији, на основу којих се врши стратификација.

Стратификација, међутим, није увек пожељан начин да се искористе подаци у вези са популацијом. Може постојати превише потенцијалних променљивих, погодних да се на основу њих врши стратификација.

За различите анализе, као најбољи се могу показати различити одабири стратума, за неке јединице је тешко одредити ком стратуму припадају и сл. Тада се прибегава постстратификацији, или тзв. "стратификацији након одабира узорка" (stratification after selection)

Постстратификација (Poststratification)

Заправо, постстратификација се може схватити као најједноставнија техника за подешавање "тежина" узорковања (adjusting sampling weights). Ради се о извесном пондерисању резултата истраживања како би се обезбедило да узорак што тачније одражава карактеристике популације из које је извучен и за коју ће се, на основу њега, доносити закључци.

Опис поступка:
Из популације изабран је СУ без понављања обима n. Читава популација подељена је, према одређеном фактору (то је помоћно обележје – auxiliary variable, категоричког типа), на J дисјунктних група - **постстратума**. Приликом узорковања, за сваку јединицу која је одабрана у узорак забележена је и њена, реализована вредност помоћног обележја, чиме је постигнуто да се свака узоркована јединица може сврстати у један од постстратума. При томе, сматра се да је унапред познат број јединица N_l у сваком постстратуму, $l = 1, 2, ..., J$.

Постстратификација (Poststratification)

Ознаке су аналогне ознакама код стратификованог узорка.

- y_{li} - вредност обележја y i-те јединице одабране у узорак у l-том постстратуму
- $\bar{y}_l$ - узорачка средина обележја за l-ти постстратум

Оцена средине представља тежинску средину узорачких средина постстратума

$$\hat{\bar{Y}}_{post} = \frac{1}{N}\sum_{l=1}^{J}\frac{N_l}{n_l}\sum_{i=1}^{n_l} y_{li} = \frac{1}{N}\sum_{l=1}^{J} N_l \bar{y}_l.$$

Ова оцена је нерпистрасна под условом да се у сваком од постстратума налази бар једна јединица из узорка.

Концептуална разлика у односу на стратификован узорак састоји се у томе што за разлику од стратификованог узорка, овде величине узорака по постстратумима n_l, нису детерминисане, него су то случајне величине.

Као сасвим добра оцена дисперзије ове оцене, за велике узорке, може се искористити следећа апроксимативна формула

$$v[\hat{\bar{Y}}_{post}] \approx \frac{N-n}{Nn} \sum_{l=1}^{J} \frac{N_l}{N} s_l^2 + \frac{N-n}{Nn^2} \sum_{l=1}^{J} \left(1 - \frac{N_l}{N}\right) s_l^2.$$

Први члан у претходном изразу, уствари је дисперзија оцене средине обележја код стратификованог случајног узорка без враћања, при пропорционалном распореду.

Други члан одражава поменуту неизвесност у вези са обимом узорака по постстратумима.

Очигледно је да је важно одабрати постстратуме који ће бити интерно што хомогенији у односу на главно обележје.

Пакет survey

Функција `postStratify(design, strata, population,...)` креира постстратификован `survey design object`.

Ту не само да су подешене "тежине" узорковања, него су и додате информације које омогућавају кориговање стандардних грешака оцена. Први корак састоји се у задавању информација о величинама подгрупа популације, тј. постстратума. Ове информације могу бити смештене у базу података или објекат типа табела (креиран функцијом table()). Ако се ради са базом података онда би она у једној (или више) својих колона требало да чува вредности променљиве (променљивих) на основу које (којих) је извршено груписање, а у последњој колокни, обавезно названој `Freq`, апсолутне фреквенције јединица популације по постстратумима.

Пакет survey

```
> data(api)
> dclus1<-svydesign(id=~dnum, weights=~pw, data=apiclus1, fpc=~fpc)
> rclus1<-as.svrepdesign(dclus1)
> pop.types <- data.frame(stype=c(''E'',''H'',''M''), Freq=c(4421,755,1018))
# ili: pop.types <- xtabs(~stype, data=apipop)
# ili: pop.types <- table(stype=apipop$stype)
> rclus1p<-postStratify(rclus1, ~stype, pop.types)
> summary(rclus1p)

Call: postStratify(rclus1, ~stype, pop.types)
Unstratified cluster jacknife (JK1) with 15 replicates.
Variables:
 [1] "cds"      "stype"    "name"     "sname"    "snum"     "dname"    "dnum"     "cname"
 [9] "cnum"     "flag"     "pcttest"  "api00"    "api99"    "target"   "growth"   "sch.wide"
[17] "comp.imp" "both"     "awards"   "meals"    "ell"      "yr.rnd"   "mobility" "acs.k3"
[25] "acs.46"   "acs.core" "pct.resp" "not.hsg"  "hsg"      "some.col" "col.grad" "grad.sch"
[33] "avg.ed"   "full"     "emer"     "enroll"   "api.stu"  "fpc"      "pw"

> svymean(~api00, rclus1p)
        mean     SE
api00 642.31 26.934
> svytotal(~enroll, rclus1p)
         total     SE
enroll 3680893 473431
```

Двофазни узорак (Two-faze Sampling)

Многе методе у теорији узорака зависе од информација о помоћној променљивој x, које су унапред прикупљене.

Када такве информације недостају, у неким ситуацијама погодно је да се на довољно великом узорку, који је извучен у првој фази узорковања, посматрају вредности само помоћне променљиве x и да се оцене њене карактеристике (средина, расподела и сл).

Оцењивање непознатих параметара у вези са "главним" обележјем y, може се, затим, урадити на узорку који се бира у другој фази, обично као подузорак узорка изабраног у првој фази, и који, јасно, садржи мањи број јединица.

Стратификован двофазни узорак

Стратификован двофазни узорак може се користити када вредности (помоћне) променљиве, које је драгоцена као критеријум за стратификацију, нису доступне за све јединице у популацији, али се релативно јефтино могу измерити.

Стратегија би била следећа:
одабрати већи узорак из популације, измерити вредности помоћне променљиве x, а онда изабрати стратификован подузорак.

Узорак одабран у првој фази може бити СУ без понављања или стратификован узорак, при чему је стратификација извршена у односу на неку другу променљиву, чије су вредности доступне на целој популацији.

Ако је узорак одабран у првој фази довољно велики, расподела вредности помоћне променљиве x на том узорку ће бити врло слична расподели њених вредности на читавој популацији, и овај план даће приближно исте оцене као стратификован једнофазни план узорковања, који имитира.

Из популације обима N прво се бира СУ без понављања обима n'.

- Ако су познате тежине (пост)стратума $\frac{N_l}{N}$ - постстратификација.
- Ако нису познате тежине (пост)стратума требало би их оценити. За почетак јединице из тог, иницијалног узорка класификују се у J (пост)стратума, и са n' означен је број јединица које су се нашле у l-том (пост)стратуму, тако је $n' = n'_1 + ... + n'_J$. Тежине стратума $\frac{N_l}{N}$ могу се оценити са n'_l, $l = 1, 2, ..., J$.

Други узорак се, потом, бира као стратификован случајан узорак без понављања из првобитног узорка, тј. из l-тог стратума се од n_l јединица бира њих n_l. За јединице одабране у узорак, у другој фази, региструју вредности посматраног обележја y.

У презентацији коришћени:

- Љ. Петровић, *Теорија узорака и планирање експеримената*, Економски факултет, Универзитет у Београду, 2007
- S. Sampath, *Sampling Theory and Methods*, Alpha Science International Ltd., Harrow, U.K. 2005
- http://r-survey.r-forge.r-project.org/survey/html/postStratify.html