ТЕОРИЈА УЗОРАКА 4

26. 04. ’13.

ДИСПЕРЗИЈА ОБЕЛЕЖЈА ПОПУЛАЦИЈЕ

- Поправљена дисперзија обележја популације дата је са

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}[Y_i - \bar{Y}]^2 = \frac{1}{N-1}\sum_{h=1}^{L}\sum_{j=1}^{N_h}[Y_{hj} - \bar{Y}]^2$$

Даље је

$$S_y^2 = \frac{1}{N-1}\sum_{h=1}^{L}\sum_{j=1}^{N_h}[(Y_{hj} - \bar{Y}_h) - (\bar{Y} - \bar{Y}_h)]^2$$

$$= \frac{1}{N-1}\left[\sum_{h=1}^{L}\sum_{j=1}^{N_h}(Y_{hj} - \bar{Y}_h)^2 + \sum_{h=1}^{L} N_h(\bar{Y} - \bar{Y}_h)^2\right]$$

$$= \frac{1}{N-1}\sum_{h=1}^{L}(N_h - 1)S_h^2 + \frac{1}{N-1}\sum_{h=1}^{L} N_h(\bar{Y} - \bar{Y}_h)^2$$

- Последњи израз може се написати и у облику

$$S_y^2 = S_W^2 + S_B^2$$

где је S_W^2 дисперзија унутар (within) стратума, а S_B^2 дисперзија између (between) стратума.

- Ако је дисперзија унутар стратума мала, ефикасност стратификације је већа.
- Поређење квалитета оцена код стратификованог случајног узорка са оценама код SRS, за исти, унапред одређен и фиксиран, обим узорка n:

Уведу се ознаке за дисперзије оцена - $V_{srs}, V_{prop}, V_{opt}$

и претпостави се да се фактор корекције популације и фактор корекције за стратуме могу занемарити. Тада важи:

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

- Нпр. ако су $V_{srs}, V_{prop}, V_{opt}$ ознаке дисперзија одговарајућих оцена средине обележја популације, под наведеним претпоставкама важи:

$$V_{srs} = \frac{S_y^2}{N} \quad V_{prop} = \frac{1}{nN}\sum_{h=1}^{L} N_h S_h^2 \quad V_{opt} = \frac{1}{nN^2}\left(\sum_{h=1}^{L} N_h S_h\right)^2$$

где је, код Neyman-овог распореда $n_h \approx N_h S_h$.

Тада је:

$$V_{srs} = V_{prop} + \frac{1}{nN}\sum_{h=1}^{L} N_h (\bar{Y} - \bar{Y}_h)^2$$

$$V_{prop} = V_{opt} + \frac{1}{nN}\sum_{h=1}^{L} N_h \left(S_h - \frac{1}{N}\sum_{h=1}^{L} N_h S_h\right)^2$$

Може се закључити да постоје две компоненте дисперзије које опадају при преласку са SRS на оптималан распоред. Прва компонента потиче од елиминације разлика између средина стратума; друга компонента потиче од елиминације ефекта разлика између стандардних одступања стратума.

ОДРЕЂИВАЊЕ ОБИМА УЗОРКА

- Нека је d апсолутна грешка оцене непознатог параметра (тј. оцене тотала или средине обележја популације), и α праг значајности.
- Код стратификованог случајног узорка потребан обим узорка за оцену тотала обележја популације је:
  - код равномерног распореда (ту је $n_h = \frac{n}{L}$)

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

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

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

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

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

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. инхерентног одступања у вредностима обележја јединица унутар истог стратума.
- Велики број стратума отежава рад и знатно повећава трошкове истраживања.

- Јасно је (између осталог, и из формула за дисперзију) да би стратуме требало формирати тако да имају што већу "хомогеност", тј. тако да је S_h^2 у сваком стратуму, $h = 1, 2, \ldots, L$, што мање. Самим тим и мали обим узорка по стратуму обезбеђује довољну прецизност оцена.
- 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/обележја значи и хомогеност вредности посматраног обележја.

У R-у

- API (пакет под називом survey) - индекс академског успеха у Калифорнији; рачунат је на основу стандардизованих тестова, које су решавали ученици калифорнијских школа. Поред података о академским постигнућима ученика по школама, на располагању су и вредности бројних 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).

Description

The Academic Performance Index is computed for all California schools based on standardised testing of students. The data sets contain information for all schools with at least 100 students and for various probability samples of the data.

Format

The full population data in `apipop` are a data frame with 6194 observations on the following 37 variables.

cds Unique identifier

stype Elementary/Middle/High School

name School name (15 characters)

sname School name (40 characters)

snum School number

dname District name

dnum District number

cname County name

cnum County number

flag reason for missing data

pcttest percentage of students tested

api00 API in 2000

api99 API in 1999

target target for change in API

growth Change in API

sch.wide Met school-wide growth target?

comp.imp Met Comparable Improvement target

both Met both targets

awards Eligible for awards program

meals Percentage of students eligible for subsidized meals

ell 'English Language Learners' (percent)

yr.rnd Year-round school

mobility percentage of students for whom this is the first year at the school

acs.k3 average class size years K-3

acs.46 average class size years 4-6

acs.core Number of core academic courses

pct.resp percent where parental education level is known

not.hsg percent parents not high-school graduates

hsg percent parents who are high-school graduates

some.col percent parents with some college

col.grad percent parents with college degree

grad.sch percent parents with postgraduate education

avg.ed average parental education level

full percent fully qualified teachers

emer percent teachers with emergency qualifications

enroll number of students enrolled

api.stu number of students tested.

The other data sets contain additional variables `pw` for sampling weights and `fpc` to compute finite population corrections to variance.

Details

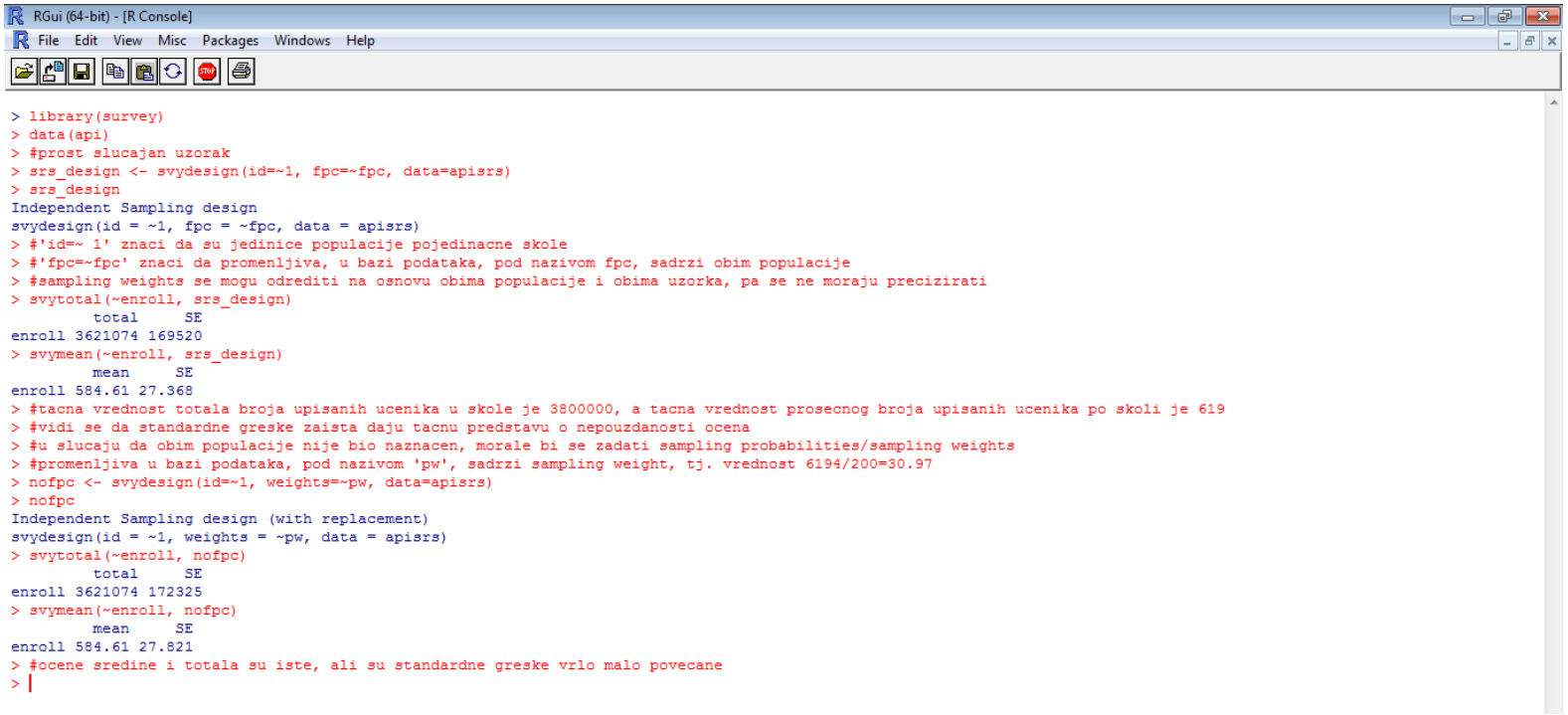
`apipop` is the entire population, `apisrs` is a simple random sample, `apiclus1` is a cluster sample of school districts, `apistrat` is a sample stratified by `stype`, and `apiclus2` is a two-stage cluster sample of schools within districts. The sampling weights in `apiclus1` are incorrect (the weight should be 757/15) but are as obtained from UCLA.

Source

Data were obtained from the survey sampling help pages of UCLA Academic Technology Services, at `http://www.ats.ucla.edu/stat/stata/Library/svy_survey.htm`.

References

The API program and original data files are at `http://api.cde.ca.gov/`



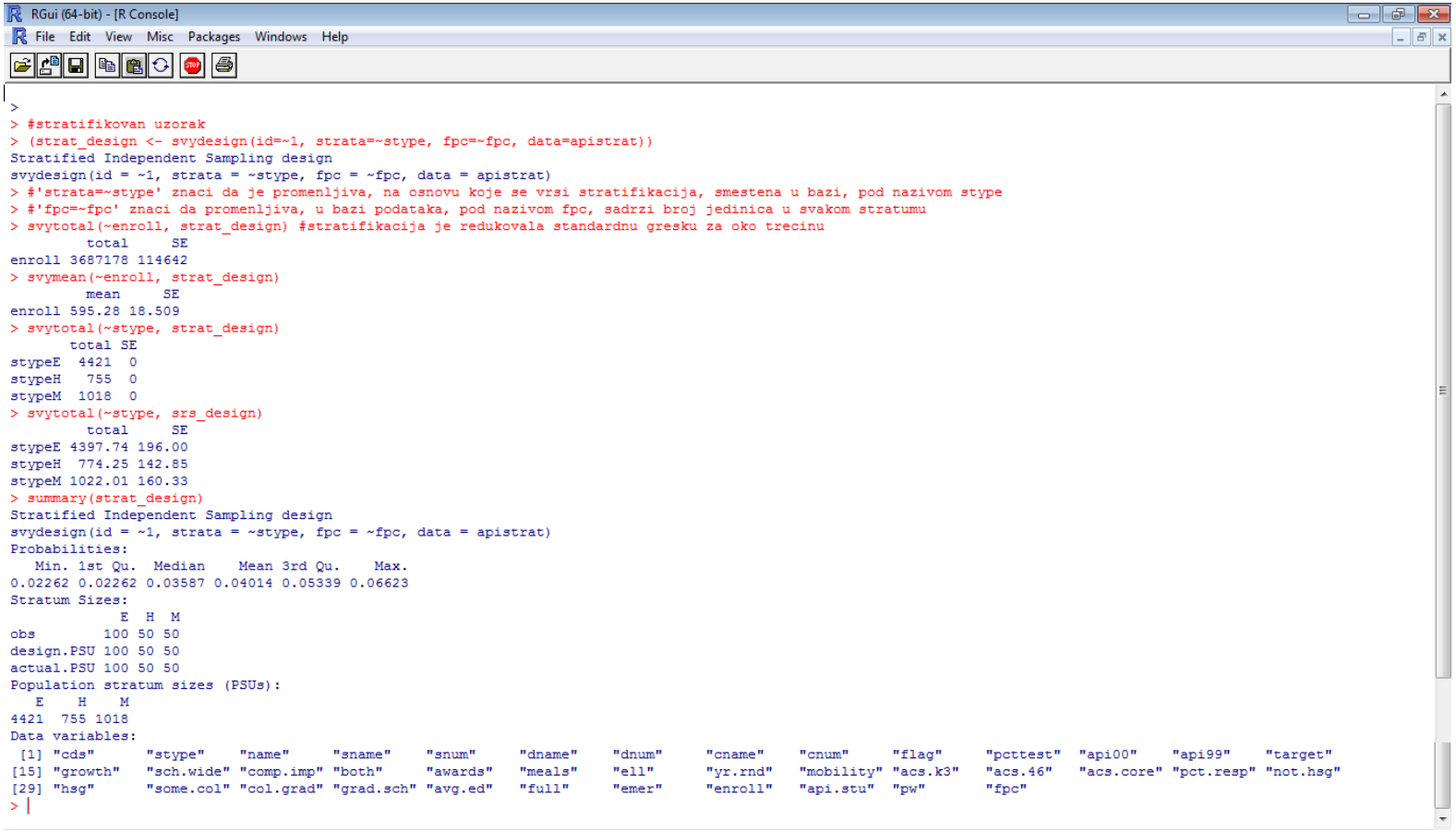
- Sampling weights

Ако се одабере прост случајан узорак од 3500 хиљаде људи из Калифорније (која има популацију од 35 милиона људи), онда свака особа има вероватноћу укључења у узорак $\pi_i = 0.0001$.

Дакле, сваки човек који се узоркује, репрезентује10000 својих сународника.

Фундаментална статистичка идеја, у позадини закључивања на основу било ког плана узорковања, јесте да јединица узоркована са вероватноћом укључења π_i репрезентује $1/\pi_i$ јединица популације. Вредност $1/\pi_i$ назива се sampling weight.

- Стратификован случајан узорак обима 200 школа, из API популације, смештен је у базу података apistrat. Стратификација је вршена на основу нивоа школовања, тј. $n_E = 100$ elementary schools, $n_M = 50$ middle schools, $n_H = 50$ high schools. Распоред узорака је направљен на основу следеће идеје: како су у Калифорнији high schools обично веће од middle schools или elementary schools, ако би се десило да SRS садржи више high schools то би водило ка "прецењеној" средини и тоталу броја уписаних ученика, док ако би се десило да SRS садржи мање high schools то би водило ка "потцењеној" средини и тоталу броја уписаних ученика. Фиксирање броја школа које треба одабрати из сваког нивоа би требало да утиче на смањење дисперзије.



- strata() – пакет под називом sampling
- пакет под називом stratification – садржи функције за једнофакторску стратификацију (univariate stratification).

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).

- Пример:

  Спроводи се испитивање јавног мњења, у форми анкете, и истраживач сматра да би стратификацију требало извршити на основу карактеристике - пол испитаника. Ако се анкета спроводи узорковањем телефонских бројева, испитаници не могу априори бити сврстани у женски, односно мушки стратум, све док не буду контактирани.

- Постстратификација је могућа онда када је поред обима популације N, познат и број јединица у сваком стратуму N_h, $h = 1, 2, \dots, L$. За разлику од стратификованог узорка, овде величине узорака по стратумима n_h, нису детерминисане, него су то случајне величине.

- Код постстратификације простог случајног узорка обима n, обим узорка n_h у стратуму h, $h = 1, 2, \dots, L$ има математичко очекивање nN_h/N, тако да добијени узорак има приближно пропорционалан распоред.

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
Double Sampling for Stratification

- 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 прост случајан узорак или стратификован узорак, при чему је стратификација извршена у односу на неку (другу помоћну) 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 прост случајан узорак обима n' .
- постстратификација - ако су познате тежине стратума (N_h/N)
- ако нису познате тежине стратума требало би их оценити. За почетак јединице из тог, иницијалног узорка се класификују у L стратума, и са n'_h је означен број јединица које су се нашле у h-том стратуму; тако је $n' = n'_1 + n'_2 + \cdots + n'_L$. Тежине стратума N_h/N се могу оценити са n'_h/n' , $h = 1, 2, \ldots, L$

  Други узорак се, потом, бира као стратификован случајан узорак из првобитног узорка, тј. из h -тог стратума се од n'_h јединица бира њих n_h . За јединице одабране у узорак, у другој фази, бележе се / мере / региструју вредности посматраног обележја y.

ДОДАТАК

Example of stratified sampling with the library `survey`:

```
# SYNTHETIC DATA
mything <- rbind(matrix(rep("nc",530),530,1,byrow=TRUE),
matrix(rep("sc",270),270,1,byrow=TRUE))
mything <- cbind.data.frame(mything,c(rep(1,350),rep(2,150),
rep(3,50),rep(1,100),rep(2,150)),rnorm(800,100,10))

names(mything) <- c("state","region","income")
table(mything$region)

n_1 <- table(mything$region)[[1]]
n_2 <- table(mything$region)[[2]]
n_3 <- table(mything$region)[[3]]

library(sampling)
s <- strata(mything,"region", size=c(50,30,20), method="srswor")
strat_thing <- getdata(mything,s)  # I extract the observed data
strat_thing$popsize <- with(strat_thing,
ifelse(region=="1",n_1,ifelse(region=="2",n_2,n_3)))

strat_thing$myweights <- 1/strat_thing$Prob

# Export data to Stata format
library(foreign)
write.dta(strat_thing,"C:/QM/mydatastrat.dta")

library(survey)
dstrata <- svydesign(id=~1, weights=~myweights, fpc=~popsize,
strat=~region, data=strat_thing)
summary(dstrata)

svymean(~income, dstrata, na.rm=TRUE)
```

```
# means by strata
svyby(~income, ~region, dstrata, svymean)

svytotal(~income, dstrata, na.rm=TRUE)
svyvar(~income, dstrata, na.rm=TRUE)

svyquantile(~income, quantile=c(0.25,0.5,0.75), design=dstrata,
na.rm=TRUE, ci=TRUE)
svyby(~income, ~region, dstrata, svyquantile, quantiles=0.5, ci=TRUE)

X11()
svyhist(~income, dstrata, main="Sample",col="pink")
X11()
svyboxplot(income ~ as.factor(region), dstrata, col="peachpuff")
X11()
plot(svysmooth(~income, design=dstrata))
```