# ТЕОРИЈА УЗОРАКА час 8

26. април ’14.

# СТРАТИФИКОВАН УЗОРАК
## НАСТАВАК

- Оцена популацијске пропорције:

  Овде се претпоставља 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` (у оквиру пакета `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 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*).

`api` *Student performance in California schools*

## 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.

### Usage

```
data(api)
```

### 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/`

Стратификован случајан узорак без понављања обима 200 школа, из `API` 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. Фиксирање броја школа које треба одабрати из сваког нивоа требало би да утиче на смањење дисперзије.

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*). Конкретно у наведеном коду овај аргумент могао је слободно бити изостављен јер се тежине могу израчунати и директно из обима стратума ($N_i/n_i$).

```
> library(survey)
> data(api)
> 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)
```

```
> 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"
```

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()`. Свака од ових функција враћа објекат са `coef`, `vcov`, `SE`, `cv` методима.

```
> svytotal(~enroll, dstrat)
          total     SE
enroll 3687178 114642
> (m <- svymean(~enroll, dstrat, deff=T))
           mean      SE  DEff
enroll 595.282  18.509 0.362
> SE(m)
         enroll
enroll 18.50851
```

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
```

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:

$$(1 \text{ x } 2000.mean) + (-1 \text{ x } 1999.mean)$$

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

Функцијом `update()` креирају се додатне променљиве унутар *survey design object*-а.

```
> dsrs1 <- update(dsrs, apidiff=api00-api99)
> dsrs1 <- update(dsrs1, apipct=apidiff/api99)
> svymean(~apidiff+apipct, dsrs1)
              mean     SE
apidiff 31.900000 2.0905
apipct   0.056087 0.0041
```

- *Sampling weights*

  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 узоркује, репрезентује $10000$ својих сународника.

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

- пакет `sampling`
  - функција `strata()`

  Служи за избор стратификованог узорка са једнаким, односно неједнаким вероватноћама.
  - функција `inclusionprobastrata()`

  Израчунава вероватноће укључења првог реда код стратификованог плана узорковања. Вероватноће укључења једнаке су за све јединице унутар истог стратума.
  - функција `Htstrata()`
- пакет `stratification`

  Садржи функције за једнофакторску стратификацију (*univariate stratification*).
- функција `epi.stratasize()`, пакет `epiR`

  Служи за добијање процене потребног обима стратификованог случајног узорка без понављања.
- ...

# 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*).

**Contracted Teaching**

| | |
|---|---|
| **Conditional Teacher Certificate:** | The conditional certificate gives a school district the flexibility to hire someone who has ***expertise*** in an area, usually when they cannot find a certificated teacher in a specific endorsement area. The certificate is subject to specific limitations and the teacher must take professional development coursework to enhance their teaching competencies. It is valid for up to 2 years.<br>**Download:** Application |
| **Emergency Teacher Certificate:** | The Emergency Certificate qualifies a teacher candidate for employment if the candidate has the appropriate degree and has ***substantially*** completed a teacher preparation program, but has ***not yet*** qualified for the Residency Certificate, if the school cannot find a regularly certificated teacher. The Emergency Certificate enables the teacher to be assigned for up to one school year.<br>**Download**: Application |
| **Transitional Teaching Certificate:** | The Transitional Teaching Certificate enables a teacher whose continuing teacher certificate has lapsed to teach for 2 years while working on reinstatement of the continuing certificate. It is valid for 2 calendar years only (24 months from issue) and can only be issued once in a teacher's career. Candidates must have a district's approval to request this certificate.<br>**Download**: Application |

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*).

- 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:

  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, $l = 1, 2, \dots, J$.

- Ознаке:

  аналогне су 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:

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

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.

- функција `postStratify()`, пакет `survey`

  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.

```
> post.st <- data.frame(stype=c("E", "M", "H"), Freq=c(4421, 1018, 755))
> dps <- postStratify(dsrs, strata=~stype, population=post.st)
> svytotal(~enroll, dps)
          total     SE
enroll 3605259 122264
```

У самом позиву функције `postStratify()` аргументом `strata` прецизиране су променљиве по којима је извршено груписање (у облику модел формуле), а аргументом `population` табела апсолутних фреквенција.

```
> # Post-stratification example
>
> # A survey of 20 x 1 m2 plots was taken from a study area of 100 m2.
> # In each quadrat the number of grubs was measured. At the same time,
> # a post-stratification into high and low quality habitats was done.
> # It was subsequently determined that there are 30 m2 of high quality
> # and 70 m2 of low quality habitat in the study area.
> grubs <- read.csv("C:/posao_fax/TU/VEZBE_1314/CAS8/post-stratify.csv", header=TRUE)
> grubs
   Grubs Post.strata
1     10           h
2      2           l
3      3           l
4      8           h
5      1           l
6      3           l
7     11           h
8      2           l
9      2           l
10    11           h
11    17           h
12     1           l
13     0           l
14    11           h
15    15           h
16     2           l
17     2           l
18     4           l
19     2           l
20     1           l
```

```
> #************ Analysis using standard R functions ***********
> # These typically do not include the effects of the finite population
> # correction factor, but will be good enough.
> #
>
> # Use the t.test function to do the same but we need to compute the se
> # based on the statistic and the estimate
> t.test.Grubs <- t.test(grubs$Grubs)
> t.test.Grubs$se.mean <- t.test.Grubs$estimate / t.test.Grubs$statistic
> cat("Est Mean Grubs per square metre is ", t.test.Grubs$estimate,
+     ";\n    se of mean(Grubs per square metre) is ", t.test.Grubs$se.mean,"\n\n")
Est Mean Grubs per square metre is  5.4 ;
    se of mean(Grubs per square metre) is  1.168445
> library(doBy)
> summary.Grubs <- summaryBy(Grubs ~ Post.strata, data=grubs,
+                            FUN=c(length,mean,sd))
> summary.Grubs$se.mean <- summary.Grubs$Grubs.sd /
+                          sqrt(summary.Grubs$Grubs.length)
> summary.Grubs
  Post.strata Grubs.length Grubs.mean Grubs.sd   se.mean
1           h            7  11.857143 3.078342 1.1635040
2           l           13   1.923077 1.037749 0.2878198
> ExpFactor <- data.frame(Post.strata=c('h','l'), area=c(30,70))
> summary.Grubs <- merge(summary.Grubs, ExpFactor)
> summary.Grubs$Grubs.total <- summary.Grubs$Grubs.mean * summary.Grubs$area
> summary.Grubs$se.Grubs.total<- summary.Grubs$se.mean  * summary.Grubs$area
> summary.Grubs
  Post.strata Grubs.length Grubs.mean Grubs.sd   se.mean area Grubs.total
1           h            7  11.857143 3.078342 1.1635040   30    355.7143
2           l           13   1.923077 1.037749 0.2878198   70    134.6154
  se.Grubs.total
1       34.90512
2       20.14739
> cat('Estimate total grubs ',sum(summary.Grubs$Grubs.total),
+     ';\n with a se of ',sqrt(sum(summary.Grubs$se.Grubs.total**2)),"\n")
Estimate total grubs  490.3297 ;
 with a se of 40.30241
```

# 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 СУ без понављања 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 СУ без понављања обима $n'$. Следи:

- постстратификација – ако су познате тежине (пост)стратума $N_l/N$

- 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 $n'_l$ означен је број јединица које су се нашле у $l$-том (пост)стратуму; тако је $n' = n'_1 + n'_2 + \cdots + n'_J$. Тежине стратума $N_l/N$ могу се оценити са $n'_l/n', l = 1, 2, \ldots, J$ .

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, бележе се / мере / региструју вредности посматраног обележја $y$.