# Групни (класер) узорак (Cluster Sampling)

8. час

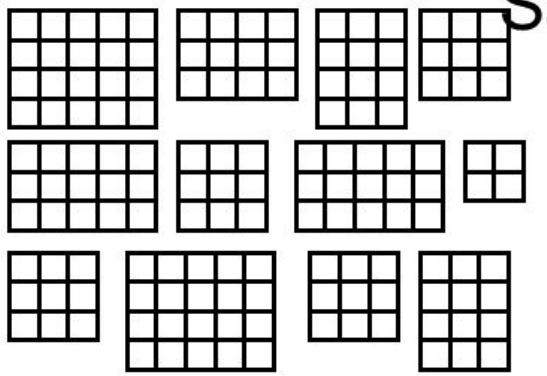
15. април 2016.

Код свих до сада описаних планова узорковања јединице су у узорак биране из целе или из стратума, на које је претходно подељена популација. Ово је погодно за истраживања малих размера, али не и за велика и комплексна истраживања. Главни разлог је тај што код популација великог обима обично не постоји употребљив списак свих јединица у популацији, на основу кога би се могао добити прост случајан, систематски или стратификован узорак. Чак и када постоји комплетан списак јединица, поменуте технике узорковања нису за примену у популацијама великог обима.

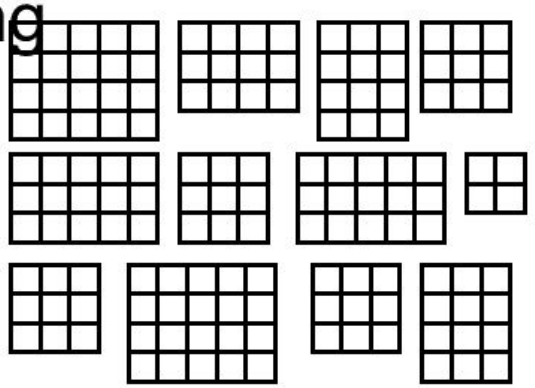
Идеја се састоји у следећем:
Популација се подели, по неком принципу, на више дисјунктних делова, који се називају примарне јединице (групе, скупине, серије, кластери), свака примарна јединица састоји се од секундарних јединица (то су, овде, јединице популације). Затим се неким вероватносним планом узорковања одабере одређени број група из којих се посматрају СВИ елементи и тако формира групни узорак.

Кластери показују површну сличност са стратумима јер кластер, као и стратум, представља групу елемената популације.
Међутим, суштинска разлика између групног и стратификованог узорка очигледна је у самом поступку селекције, тј. одабира јединица.

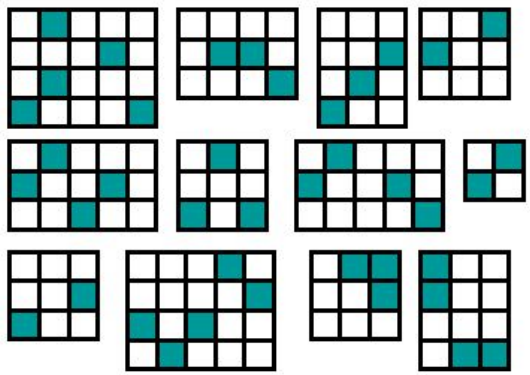
## Difference Between Cluster and Stratified Sampling



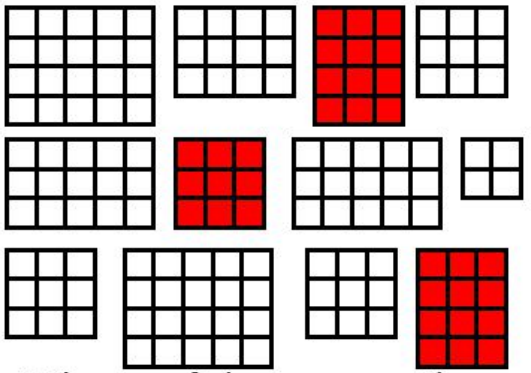
Population of $L$ strata, stratum $l$ contains $n_l$ units



Population of $C$ clusters



Take simple random sample in *every* stratum



Take srs of clusters, sample every unit in chosen clusters

Да би оцене биле прецизне потребно је да кластери по својој структури што боље одсликавају популацију. У томе се састоји и њихова главна разлика у односу на стратуме, који се формирају као интерно хомогене групе јединица.

Стратификован узорак се користи за доношење што прецизније оцене непознатих параметара, док се групни узорак користи када је потребно смањити трошкове истраживања.

Са друге стране, испитивањем свих јединица кластера, делимично се понавља иста информација, тј. добија се мање нових информација него што је случај код простог случајног узорка.

Према томе, групни узорак је мање прецизан од стратификованог и простог случајног узорка.

Претпостави се да је популација подељена на $L$ кластера, при чему $i$-ти кластер садржи $N_i$ јединица, $i = 1, 2, ..., L$.

Нека је $Y_{ij}$, $j = 1, 2, ..., N_l$, $i = 1, 2, ..., L$, вредност обележја $y$ $j$-те јединице из $i$-тог кластера, а

$$Y_i = \sum_{j=1}^{N_i} Y_{ij}$$

је тотал обележја $i$-тог кластера $i = 1, 2, ..., L$.

Даље, од $L$ кластера се врши одабир СУ без понављања $s$ обима $n$ (кластера).

Непристрасна оцена тотала $Y$ обележја популације код групног узорка је

$$\hat{Y}_{cls} = \frac{N}{n} \sum_{i \in s} Y_i.$$

Други могући приступ јесте да се од $L$ кластера не врши одабир СУ без понављања, него узорка са вероватноћама пропорционалним величини са понављањем, обима $n$.

Наиме, идеја је да се број јединица унутар сваког кластера посматра као помоћна променљива – "величина".

Вероватноћа избора $r$-тог кластера је, тада, $p_r = \frac{N_r}{N_0}$, $r = 1, 2, ..., L$, где је $N_0 = \sum_{i=1}^{L} N_i$.

Систематски узорак је, заправо, специјалан случај групног узорка, код кога се од $k$ кластера бира тачно један, као случајан узорак, а затим посматрају вредности испитиваног обележја на свим јединицама у том кластеру.

# ВИШЕЕТАПНИ УЗОРАК (Multistage Sampling)

Претпоставља се да је популација подељена на одређен број, нпр. $L$, примарних јединица. Ако се прво бира узорак одређеног обима, нпр. $n$, примарних јединица, а затим бира узорак од секундарних јединица из сваке изабране примарне јединице, такав план узорковања назива се двоетапни узорак (two-stage sampling).

Предност се даје примени двоетапног узорка у односу на групни узорак, у ситуацијама када су кластери велики (у смислу садрже велики број секундарних јединица) или када су секундарне јединице унутар кластера веома сличне, па испитивање СВИХ секундарних јединица садржаних у примарној јединици може бити скупо и непотребно.

Понављањем описаног поступка добија се вишеетапни узорак (multi-stage sampling).

# Пример троетапног узорка

Врши се анкетирање ученика средњих школа у неком граду. Прво се одабере узорак школа, затим из узорак одељења из одабраних школа, и на крају узорак ученика у одабраним одељењима.

Термин ‘примарне јединице’ резервисан је за највеће групе, тј. за делове популације из којих се избор врши у првој етапи. Наредне у хијерархији су ‘секундарне јединице’, потом ‘терцијарне јединице’ итд. У последњој етапи формира се узорак кога чине ‘праве јединице’ популације.

# Пример - API популација

База података `apiclus2` садржи двоетапни групни узорак који се састоји од $40$ школских округа и највише по $5$ школа, узоркованих из сваког.

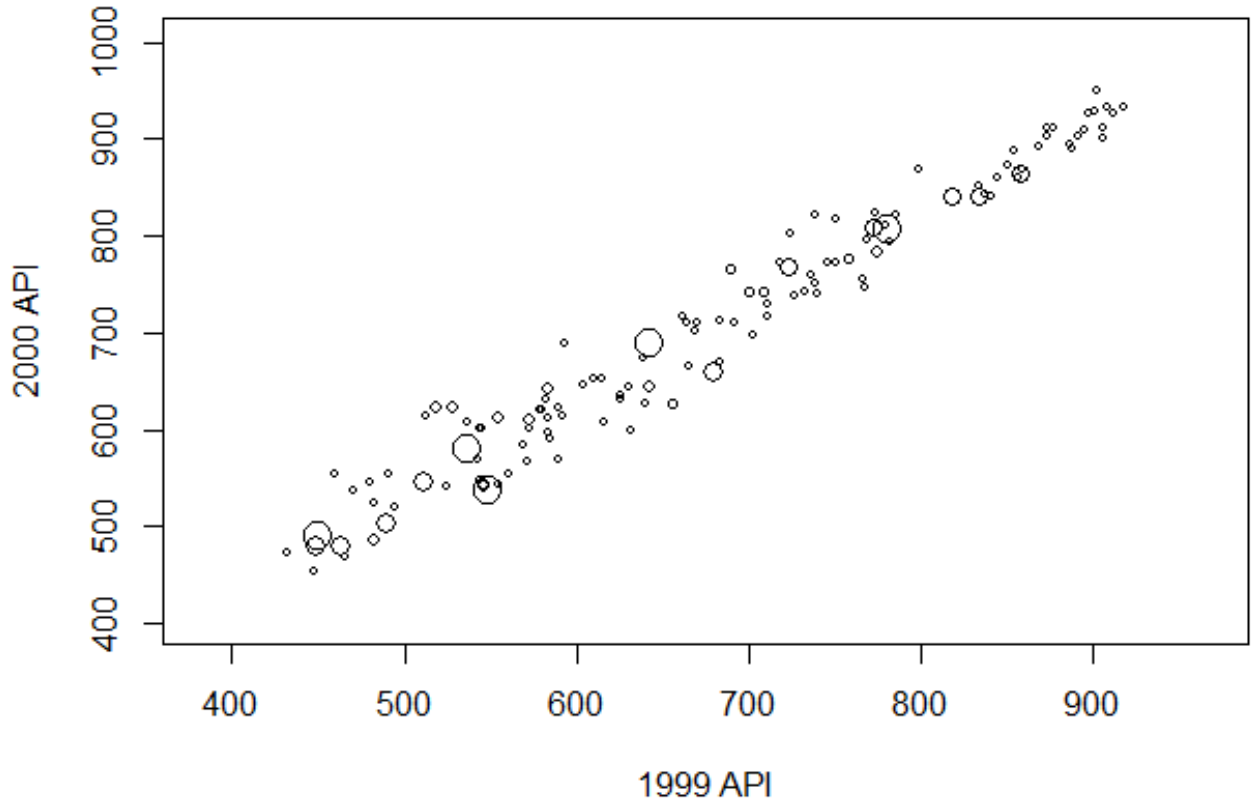
```
> dclus2 <- svydesign(id=~dnum+snum, fpc=~fpc1+fpc2, data=apiclus2)
> summary(dclus2)
```

```
2 - level Cluster Sampling
With (40, 126) clusters.
svydesign(id = ~dnum + snum, fpc = ~fpc1 + fpc2, data = apiclus2)
Probabilities:
 Min. 1st Qu. Median Mean 3rd Qu. Max.
0.003669 0.037740 0.052840 0.042390 0.052840 0.052840
Population size (PSUs): 757
Data variables:
[1] "cds" "stype" "name" "sname" "snum" "dname" "dnum"
[8] "cname" "cnum" "flag" "pcttest" "api00" "api99" "target"
[15] "growth" "sch.wide" "comp.imp" "both" "awards" "meals" "ell"
[22] "yr.rnd" "mobility" "acs.k3" "acs.46" "acs.core" "pct.resp" "not.hsg"
[29] "hsg" "some.col" "col.grad" "grad.sch" "avg.ed" "full" "emer"
[36] "enroll" "api.stu" "pw" "fpc1" "fpc2"
```

У позиву функције `svydesign()`, променљивом `dnum` идентификују се школски округи, а променљивом `snum` школе, `fpc1` је број школских округа у читавој популацији, а `fpc2` број школа у округу.
"Тежине" узорковања израчунавају се помоћу `fpc1` и `fpc2`.

```
> svymean(~api00, dclus2, deff=T)
 mean SE DEff
api00 670.812 30.099 6.2505
> svyvar(~api00, dclus2)
 variance SE
api00 18722 2134
> svymean(~factor(stype), dclus2)
 mean SE
factor(stype)E 0.68118 0.0556
factor(stype)H 0.13432 0.0567
factor(stype)M 0.18450 0.0222
> svymean(~api00+meals+ell+enroll, dclus2, deff=T, na.rm=T)
 mean SE DEff
api00 673.0943 31.0574 6.2833
meals 52.1547 10.8368 11.8585
ell 26.0128 5.9533 9.4751
enroll 526.2626 80.3410 6.1427
> pop.types <- data.frame(stype=c("E", "H", "M"), Freq=c(4421, 755, 1018))
> dps <- postStratify(dclus2, strata=~stype, population=pop.types)
> svytotal(~enroll, dclus2, na.rm=T)
 total SE
enroll 2639273 799638
> svytotal(~enroll, dps, na.rm=T)
 total SE
enroll 3074076 292584
```

```
> svyplot(api00~api99, design=dclus2, style="bubble", xlab="1999 API", ylab="2000 API", inches=.1)
```



# ГРЕШКЕ УЗОРКА (Sampling Errors)

До грешке узорка долази зато што се испитивање (неког обележја, нпр. $y$) врши само на узорку, тј. на делу популације обима нпр. $n$ јединица, а не и на целој популацији обима нпр. $N$ јединица.

Грешка узорка проузрокује разлику између оцене непознатог параметра и праве вредности тотала или средине обележја популације.

Наравно, овде се претпоставља да је вредност посматраног обележја за $i$-ту јединицу $-$ $y_i$ права вредност за ту јединицу.

# ГРЕШКЕ ВАН УЗОРКА (Non-sampling Errors)

Додатне грешке могу настати у ситуацијама када се код јединица, одабраних у узорак, не забележе тачни подаци о њиховим карактеристикама, које су од интереса при истраживању, затим приликом обраде и анализе прикупљених података, при табелирању, па чак и код објављивања добијених резултата.

То су грешке ван узорка, тј. неузорачке грешке.

Оне могу бити веће од узорачких грешака, и за разлику од истих, повећање обима узорка нема ефекта на њихово смањење.

Обично су израженије у истраживањима већих размера и, у принципу, теже их је квантификовати и контролисати, него узорачке грешке.

# ЕФЕКТИ НЕПОТПУНОСТИ ПОДАТАКА (Incomplete Surveys)

Један од честих узрока грешке ван узорка јесте непотпуност података.

Наиме, у пракси, приликом спровођења истраживања, често се дешава да није могуће прикупити информације за све јединице из узорка, па подаци добијени на основу узорка нису потпуни, што ствара проблеме приликом оцењивања и утиче на квалитет добијених резултата.

Пример:
Код истраживања већих размера у људској популацији, испитаник који је одабран у узорак можда није доступан у моменту спровођења истраживања или, ако јесте, може да одбије сарадњу са истраживачем и сл.

Неки од најчешћих узрока, који доводе до непотпуности података:

- необухватност свих јединица одабраних у узорак до које може доћи из више разлога (некомплетни спискови јединица у узорку, слаба комуникација у истраживачком тиму итд)
- немогућност да се добију подаци/"одговори" од испитиване јединице
- приликом истраживања у људској популацији – неспособност испитаника да да одговоре на одређена питања, из различитих разлога (недовољна обавештеност о одређеној тематици, неписменост итд)
- приликом истраживања у људској популацији – одбијање испитаника да учествује у истраживању

# НЕВЕРОВАТНОСНО УЗОРКОВАЊЕ

Често се у истраживањима користе и узорци који нису формирани случајним избором јединица популације.

Такав избор јединица изводи се из различитих разлога и у разне сврхе.

При одлучивању о избору плана узорковања, важно је знати да узорци који нису на случајан начин добијени из популације, нису погодни за генерализацију резултата и налаза истраживања на целу популацију.

# Квотни узорак (Quota Sampling)

Потребно је прво јасно дефинисати популацију, која се, затим, дели на подскупове јединица (субпопулације) према карактеристикама које су интересантне за истраживање, а које су одређене проблемом, предметом и циљевима истраживања. Потом се одређује величина сваке субпопулације, потребна величина узорка и квоте, тј. број јединица сваке субпопулације које треба изабрати у узорак, али тако да распоред узорка буде пропорционалан (броју јединица у субпопулацијама у односу на обим целе популације).

Коначно, избор потенцијалних јединица из сваке субпопулације у узорак врши се слободним просуђивањем и одлучивањем истраживача.

Овакво узорковање не захтева велике трошкове (време, ангажовање, материјални трошкови,...), па се, због свега тога, сматра да је најзначајнији план невероватносног узорковања.

# Пригодни узорак (Availability Sampling)

Сачињавају га јединице популације које су расположиве и лако доступне истраживачу. Према томе, постоје и јединице у популацији за које не постоји никаква могућност да буду изабране у узорак. Заправо, често остаје нејасно које све јединице чине популацију. Овако формирани 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 узорка.

Пригодни узорак се може користити у експлоративним истраживањима, где генерализација резултата не долази у први план, јер су то обично почетна истраживања за серију накнадних, у којима би требало, на коректним узорцима, потврдити или одбацити почетне налазе.

# Намерни узорак (Purposive Sampling)

Јединице се из популације бирају у узорак на основу својстава, која су прецизно дефинисана проблемом, предметом и циљевима истраживања (у људској популацији нпр. на основу одређених способности, активности, жеље да учествују у истраживању, знања из области од интереса итд). Узорковање се заснива на просуђивању истраживача. Наравно, резултати истраживања се односе на јединице знатно ужег подскупа јединица популације.

Намерни узорак је погодан за експлоративна истраживања, посебно за третирање екстремних случајева у популацији.

Прецизнији је у односу на пригодни узорак.

# Узорак “снежних грудви” (Snowball Sampling)

Користи се за аналитичка истраживања, и то искључиво у људској популацији. Формира се тако што се одабере неки број испитаника, који током прикупљања података указују на нове испитанике, који би могли ући у узорак. Ти нови испитаници би могли, даље, да укажу на наредне, и поступак долажења до нових испитаника је сличан ефекту снежне грудве, те отуда и назив.

Узорак “снежних грудви” прикладан за испитивање својстава која се у популацији ретко јављају

# Хвала на пажњи! .