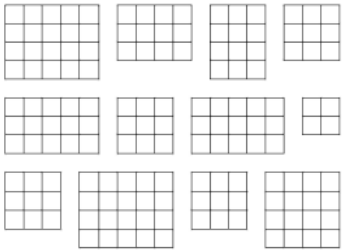
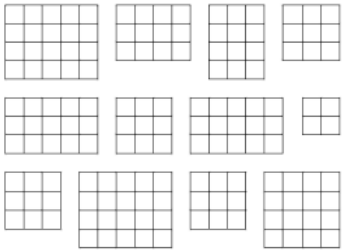
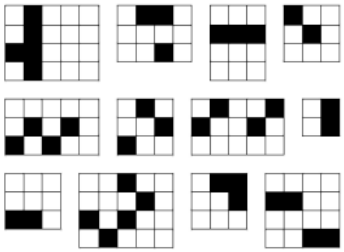
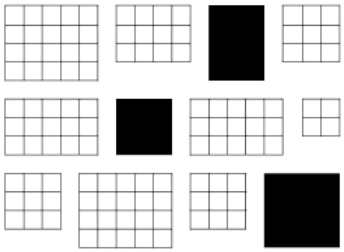
# ТЕОРИЈА УЗОРАКА час 10

7. мај ’14.

# ГРУПНИ УЗОРАК
(*Cluster Sampling*)

- Код свих до сада описаних планова узорковања јединице су у узорак биране из целе популације 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 економичне за примену у популацијама великог обима.

- *Single-Stage Cluster Sampling*
  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.

| Stratifikovano uzorkovanje | Klaster uzorkovanje |
|---|---|
| Svaki element populacije se nalazi u tačno jednom stratumu. | Svaki element populacije se nalazi u tačno jednom klasteru. |
| Populacija od $H$ stratuma: stratum $h$ ima $N_h$ elemenata  | Jednofazno klaster uzorkovanje: populacija od $N$ klastera  |
| Uzima se SRS iz svakog stratuma:  | Uzima se SRS klastera i posmatraju se svi elementi unutar izabranih klastera:  |
| Disperzija ocene srednje vrednosti populacije $\bar{y}_U$ zavisi od promenljivosti unutar stratuma. | Klaster je uzoračka jedinica. Što više klastera uzorkujemo, manja je disperzija. Disperzija ocene srednje vrednosti populacije $\bar{y}_U$ zavisi prvenstveno od promenljivosti između sredina klastera. |
| Za postizanje veće preciznosti, pojedinačni elementi unutar svakog stratuma treba da budu što sličniji, a sredine stratuma treba da se razlikuju što je više moguće. | Za postizanje veće preciznosti, pojedinačni elementi unutar svakog klastera treba da budu što različitiji, dok sredine klastera treba da budu što sličnije. |

Резимирано, 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, $i = 1, 2, \dots, L$. Нека је $Y_{ij}$, $j = 1, 2, \dots, N_i$, $i = 1, 2, \dots, L$, 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 $i = 1, 2, \dots, L$.

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

  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, $p_r = \frac{N_r}{N_0}$ $r = 1, 2, \ldots, L$ , 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.

- Пример

Црвени морски јежеви сматрају се посластицом и њихово ловљење у Британској Колумбији доноси приход од неколико милиона долара по сезони.

Да би се могао проценити улов и пратити залихе ових животиња важно је одредити густину њихове насељености сваке године. У том циљу, прво се повуку линије управне на обалу, на местима где је познато да постоје легла јежева. Затим, рониоци пливају дуж линије и котрљају PVC квадрат димензије $1\text{m} \times 1\text{m}$ , такође, дуж линије и броје колико је јединки морског јежа у том квадрату одговарајуће величине да се смеју ловити, а колико је оних чија је величина мања од дозвољене (мања од $7\text{cm}$, тј. $2.7\ \text{in}$).

**Дакле, у овом конкретном примеру популацију од интереса чине морски јежеви у области где се врши њихово прикупљање. Истраживање је вршено тако да су животиње, практично, вештачки груписане у зависности од линије на којој су узорковане. Како су тачке дуж обале, са којих су повлачене линије, одабране коришћењем случајног узорка без понављања сматра се да се у бази налазе подаци за случајан узорак кластера без понављања.**

**Параметар од интереса је густина насељености јежева дозвољене величине.**

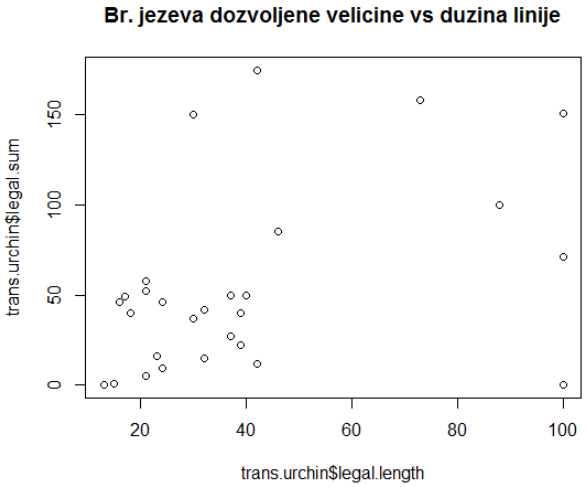
База садржи 1120 опсервација и четири променљиве: transect – редни број линије, quad – редни број квадрата у вези са одговарајућом линјиом, legal, sublegal – број јединки дозвољене величине, односно величине мање од дозвољене у одговарајућем квадрату. Када се погледа база може се уочити да недостају линије нумерисане са 5, 12, 17, 19, 32. Према евиденцији истраживања, елементарне непогоде су проузроковале отказивање испитивања ових линија. Може се сматрати да недостајуће линије овде недостају на потпуно случајан начин.

```
> urchin <- read.table("C:/posao_fax/TU/VEZBE_1314/CAS10/urchin.csv", sep=",", header=T)
> str(urchin)
'data.frame':  1120 obs. of  4 variables:
 $ transect: int  1 1 1 1 1 1 1 1 1 1 ...
 $ quad    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ legal   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ sublegal: int  0 0 0 1 0 0 0 0 0 0 ...

> urchin$total <- urchin$legal + urchin$sublegal
> #prvo bi trebalo napraviti grupni prikaz podataka po linijama
> library(doBy)
> trans.urchin1 <- summaryBy(legal ~ transect, data=urchin, FUN=c(length,sum))
> trans.urchin2 <- summaryBy(quad  ~ transect, data=urchin, FUN=c(min,max))
> trans.urchin <- merge(trans.urchin1,trans.urchin2)
> #pozvati: trans.urchin
> cat('Provera da li je izgubljena informacija o nekom kvadratu\n')
Provera da li je izgubljena informacija o nekom kvadratu
> trans.urchin$legal.length - (trans.urchin$quad.max-trans.urchin$quad.min+1)
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> #NIJE!
```

Даља идеја јесте да се искористи количничка оцена, где би главно обележје било број морских јежева дозвољене величине, мерених на одређеној линији, а помоћно обележје величина линије (тј. број испитиваних квадрата на тој линији).

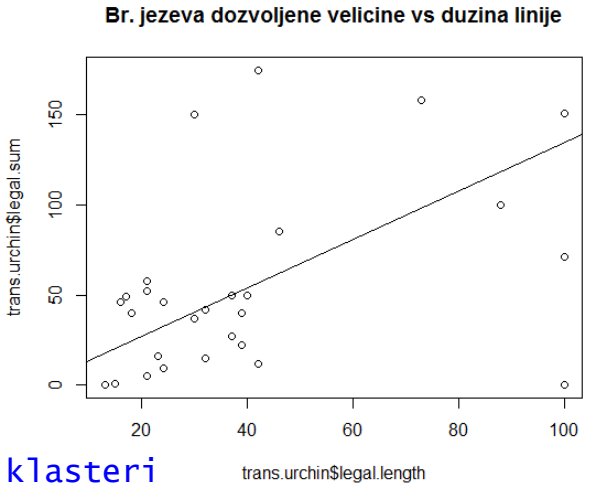
Пре тога, требало би исцртати податке и оправдати примену количничког оцењивања.

```
> #dalje, iscrtavaju se podaci kako bi videlo da li je njihova veza približno linearna
> #i da li regresiona prava prolazi kroz koordinatni pocetak
> plot( trans.urchin$legal.length, trans.urchin$legal.sum,
      main='Br. jezeva dozvoljene velicine vs duzina linije')
```




```
> #pravi se model linearne regresije sa uslovom da (buduca) regresiona prava prolazi kroz
koordinatni pocetak
> urchin.fit <- lm(legal.sum ~ 0 + legal.length, data=trans.urchin, weights=1/legal.length)
> summary(urchin.fit)

Call:
lm(formula = legal.sum ~ 0 + legal.length, data = trans.urchin,
    weights = 1/legal.length)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-13.4554  -4.8586  -0.6094   4.0849  20.0163

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
legal.length   1.3455     0.2161   6.226 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.232 on 27 degrees of freedom
Multiple R-squared:  0.5895,   Adjusted R-squared:  0.5743
F-statistic: 38.77 on 1 and 27 DF,  p-value: 1.168e-06
```

```
> plot(trans.urchin$legal.length, trans.urchin$legal.sum,
      main='Br. jezeva dozvoljene velicine vs duzina linije')
> abline(a=0, b=urchin.fit$coefficients)
```




```
> library(survey)
> urchin.design <- svydesign(data=trans.urchin,
+                            ids=~transect, #linije su grupe, tj. klasteri
+                            variables=~legal.sum+legal.length,
+                            fpc=NULL)  #nije zadat obim populacije ni tezine uzorkovanja
Warning message:
In svydesign.default(data = trans.urchin, ids = ~transect, variables = ~legal.sum +  :
  No weights or probabilities supplied, assuming equal probability
```

Обим популације (коју чине примарне јединице – линије) није задат, јер могућих линија има бесконачно много, стога се игнорише и фактор корекције због коначности популације.

```
> print(urchin.design)
Independent Sampling design (with replacement)
svydesign(data = trans.urchin, ids = ~transect, variables = ~legal.sum +
    legal.length, fpc = NULL)
> (urchin.ratio <- svyratio(numerator=~legal.sum, denominator=~legal.length, urchin.design))
Ratio estimator: svyratio.survey.design2(numerator = ~legal.sum, denominator = ~legal.length,
    urchin.design)
Ratios=
           legal.length
legal.sum     1.345536
SEs=
           legal.length
legal.sum     0.227248
```

# 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 (*multistage 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.

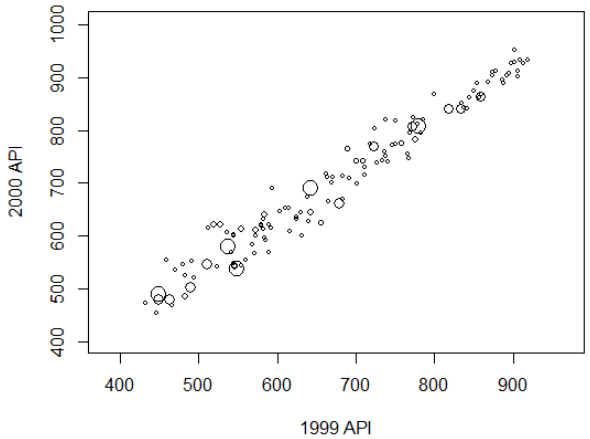
- # Пример - API популација

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

```
> dclus2 <- svydesign(id=~dnum+snum, fpc=~fpc1+fpc2, data=apiclus2)
> summary(dclus2)
2 - level Cluster Sampling design
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.

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

> #poststratifikacija
> 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
```



# 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/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. (*Non-response*).

- 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

- Ефекти непотпуности података су понекад толико изражени да потпуно деформишу (“изопачују”) резултате.
- Развијено је неколико техника за одстрањивање пристрасности, настале услед некомплетности података:
  - *Hansen & Hurwitz Technique*
  - *Deming’s Model for the effects of call-backs*
  - ...

# ЕФЕКТИ ГРЕШАКА “МЕРЕЊА” (*Observational Errors*)

- До сада је претпостављано да је вредност посматраног обележја за $i$-ту јединицу − $y_i$ права (у смислу, тачна) вредност за ту јединицу.
- Међутим, та претпоставка је често сувише поједностављена у односу на реалну ситуацију, и искуство је не подржава.
- Постоји мноштво примера, који показују да је грешка мерења при узорковању присутна, у току вршења истраживања.

# ГРЕШКЕ ПРИКУПЉАЊА ПОДАТАКА УЗОРКОМ – РЕЗИМИРАНО

Šta je greška?

- Razlika između stvarne vrednosti (u populaciji) i opservirane vrednosti (u uzorku)

Vrste grešaka?

- Uzoračke greške
- Neuzoračke greške – sve ostale greške u vezi sa istraživačkim projektom. Mogu se podeliti u četiri grupe:
  - **Greške dizajna** – usled propusta u istraživačkog dizajna (greška izbora, greška specifikacije populacije, greška u uzoračkom okviru, greška zamene informacija, greška merenja, greška dizajniranja eksperimenta, greška u obradi podataka).
  - **Greške u sprovođenju** – greške koje nastaju tokom primene anketnog instrumenta na ispitanike (greška u postavljanju pitanja, greška evidentiranja i greška uticaja anketara).
  - **Greške odgovora** – kada ispitanik pruži netačne odgovore.
  - **Greške neodgovora** – kada neki ispitanici nisu kontaktirani ili jesu kontaktirani ali pruže nepotpune odgovore / ne pruže nikakav odgovor

# 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.

# *** ДИЗАЈН И ПЛАНИРАЊЕ ЕКСПЕРИМЕНАТА ***

# ХВАЛА НА ПАЖЊИ!!!