The State Statistical Committee of the Republic of Azerbaijan

Centre of Scientific Research and Statistical Innovation

Using R in sampling and estimating

- Mubariz Nuriev- Direktor of Centre of Scientific Research and Statistical Innovation
- Zarifa Nagieva - Head of the Department of Scientific Research and Statistical Research
- Saleh Movlamov - Chief Economist of the Department of Scientific Research and Statistical Research

Household survey

- Household surveys are a key part of socio-economic statistics and a source of information on country development trends.
- Household surveys include the following steps:

  -planning of research;

  -preparation of frame for survey;

  -preparation of survey tools;

  -assessment of the observation area;

  -determination of the number of samples;

-grouping households by urban, rural settlements;

-selecting households for household surveys;

-collection and analysis microdata;

-estimate data of sample survey;

-desimination result of sample survey.

Codes for sampling proccess in R

```
#import table from file to object resp
resp<-read.csv2("C:/resp.csv",header=TRUE, sep=";")
if (file.exists(file = "C:/resp_secme.csv"))
#Delete file if it exists
file.remove(file = "C:/resp_secme.csv")
resp
resp_ray<-unique(resp$erazi_kod)
resp_ray
n<-length(resp_ray)
```

```
#Allocate sample frame by region - urban and rural area
n
i=1
while (i<=n) {
j<-resp_ray[i]
j
resp_ray_j<-resp[resp$erazi_kod==j,]
resp_ray_j
resp_ray_sh_kn<-unique(resp_ray_j$sh_kn)
resp_ray_sh_kn
q<-length(resp_ray_sh_kn)
q
h=1
while (h<=q) {
sh<-resp_ray_sh_kn[h]
sh
```

grouping of urban and rural areas

```
resp_ray_sh_kn_j<-resp_ray_j[resp_ray_j$sh_kn==sh,]
resp_ray_sh_kn_j
resp_ray_sh_kn_v4<-unique(resp_ray_sh_kn_j$v4)
resp_ray_sh_kn_v4
f_sh<-length(resp_ray_sh_kn_v4)
f_sh
s=1
    while (s<=f_sh) {
    k<-resp_ray_sh_kn_v4[s]
    k
    resp_ray_sh_kn_v4_k<-resp_ray_sh_kn_j[resp_ray_sh_kn_j$v4==k,]
    resp_ray_sh_kn_v4_k
    l<-nrow(resp_ray_sh_kn_v4_k)
    l
     m<-resp_ray_sh_kn_v4_k$sec_say
     m
```

```
#sampling process on region (urban and rural area)
sec_evj<-sample(1:nrow(resp_ray_sh_kn_v4_k), m)
sec_evj
a<-resp_ray_sh_kn_v4_k[sec_evj,]
a
# save as resulte sampling process
write.table(a, file = "C:/resp_secme.csv", append = TRUE, sep = ",",
col.names = NA,qmethod = "double")
s<-s+1
        }
h=h+1
      }
i<-i+1
}
```

Estimating in R

- Verification and editing of data of respondents is carried out at the micro and macro level in data editing process.
- The microdata refers statistical observation units and restored of suspicious and incomplete indicators by various means.
- Data macro editing is done at the regional or macro level.

The following indicators are used to check the data processing process:

N: Number of observation units;

Nc: The number of observation units that refused the request;

Nr: The number of renewable (imputation) observation units;

X: Total initial (x) indicators for all observation units;

Xc: Total initial (XAM) indicators for the observed surveillance units;

Yi: Total reversed indicators on the abandoned observation units;

A: Total number of edited indicators for all observation units;

Kc: Control expenses for the review;

Ki: Recovery costs (imputation).

Depending on the size of the sample and other parameters, the results of the sample survey is calculated from microdata and weight factor, which is defined by the following formula:

$$W_i = W_{bj} * K_{1j} * K_{2j} * K_{3j} * K_{4j} * K_{5j}$$

where: W_i - is the resulting weight for the i-respondent;

W_{bj}-is the base weight of the j-household;

K_{1j} -is the theoretical coefficienr of the household sampling probability;

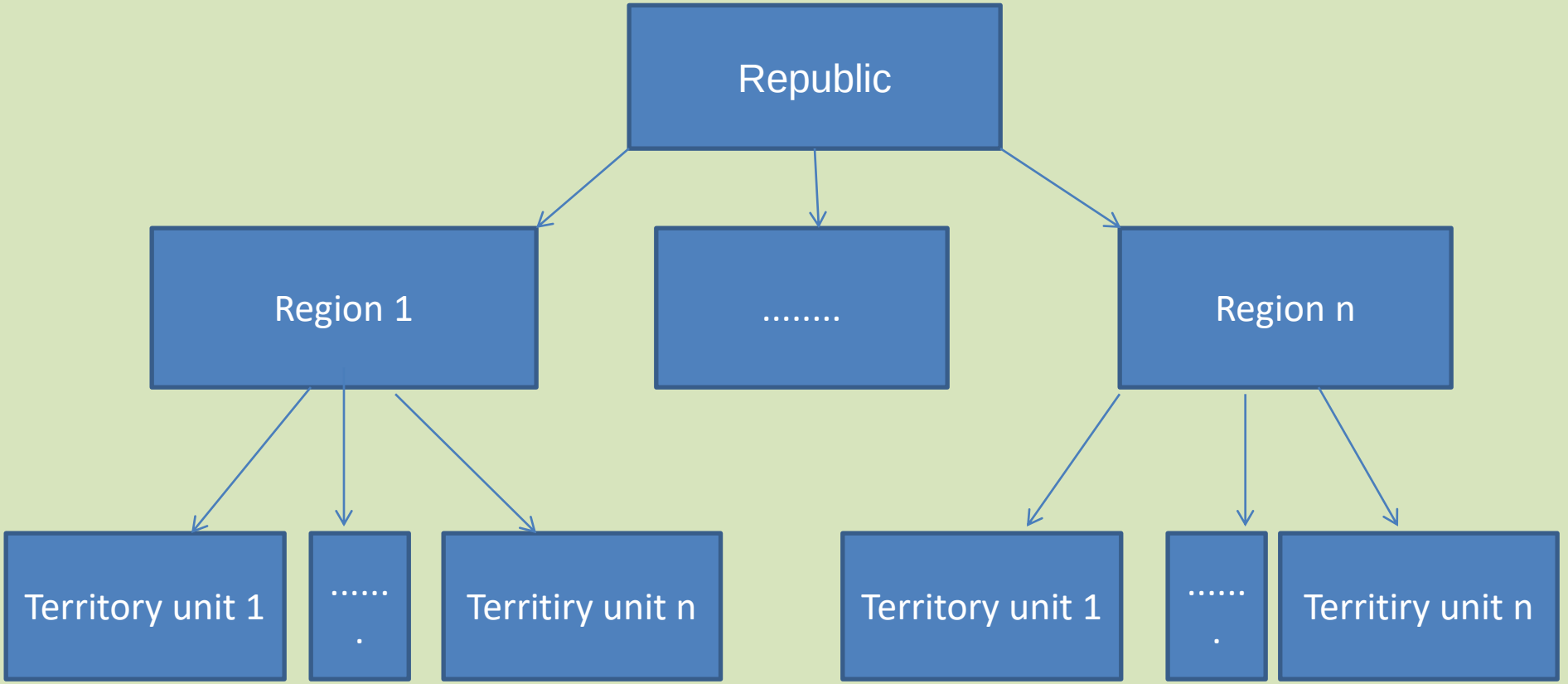
K_{2j} -is the coefficient of non-dwelling and locked premises;

K_3j- is the sampled household "non-response" coefficient;

K_{4j} - is the respondent "non-response" coefficient;

K_{5j} - is the post-stratification compensation.

As the statistical information system is in the form of a network, the scheme for the generalization of the results of research is as follows.



The codes for application in R are following:

```
# Module created in R programming system
# To estimate household spending on regions
# Census data and observation data were used
        rt=0
        rSt=0
        j=1
        # The codes by regions
         while (j<=11)
         {
         code_region<-c(0,1,2,3,4,5,6,7,8,9,10)
         i=code_region[j]
 #Reading data from household register
 #Region code, family size, number of family
 pop<-read.csv2("c:/region.csv",header=TRUE, sep=";")
```

```
#Reading data from household register
#Region code, family size, number of family
pop<-read.csv2("c:/region.csv",header=TRUE, sep=";")
pop_hh<-pop[pop$reg==j, ]
# Grouping on family size in the regions
pop_hh1<-pop_hh[pop_hh$hh_size==1, ]
pop_hh2<-pop_hh[pop_hh$hh_size==2, ]
pop_hh3<-pop_hh[pop_hh$hh_size==3, ]
pop_hh4<-pop_hh[pop_hh$hh_size==4, ]
pop_hh5<-pop_hh[pop_hh$hh_size==5, ]
pop_hh6<-pop_hh[pop_hh$hh_size==6, ]
pop_hh7<-pop_hh[pop_hh$hh_size==7, ]
```

```
#Calculating the total number of families in regions
        N1<-sum(pop_hh1$number,na.rm=T)
        N2<-sum(pop_hh2$number,na.rm=T)
        N3<-sum(pop_hh3$number,na.rm=T)
        N4<-sum(pop_hh4$number,na.rm=T)
        N5<-sum(pop_hh5$number,na.rm=T)
        N6<-sum(pop_hh6$number,na.rm=T)
        N7<-sum(pop_hh7$number,na.rm=T)
  # reading of observation data
  # table of region code, family income, family size
  sample_base<-read.csv2("c:/baza.csv",header=TRUE, sep=";")
  sample_base<-sample_b[sample_b$region==0, ]
```

```
# Grouping on family size of observation data
sample_base1<-sample_base[sample_base$hh_size==1,]
sample_base2<-sample_base[sample_base$hh_size==2,]
sample_base3<-sample_base[sample_base$hh_size==3,]
sample_base4<-sample_base[sample_base$hh_size==4,]
sample_base5<-sample_base[sample_base$hh_size==5,]
sample_base6<-sample_base[sample_base$hh_size==6,]
sample_base7<-sample_base[sample_base$hh_size==7,]
# Calculation  the number of families on observation data
n1<-nrow(sample_base1)
 n2<-nrow(sample_base2)
n3<-nrow(sample_base3)
n4<-nrow(sample_base4)
n5<-nrow(sample_base5)
n6<-nrow(sample_base6)
```

```
# Calculation the average value
                if (n1 > 0 )
{mean1<- mean(sample_base1$cons_exp,na.rm=T)}
                if (n2 > 0 )
{mean2<- mean(sample_base2$cons_exp,na.rm=T)}
                if (n3 > 0 )
{mean3<- mean(sample_base3$cons_exp,na.rm=T)}
                if (n4 > 0 )
{mean4<- mean(sample_base4$cons_exp,na.rm=T)}
                if (n5 > 0 )
{mean5<- mean(sample_base5$cons_exp,na.rm=T)}
                if (n6 > 0 )
 {mean6<- mean(sample_base6$cons_exp,na.rm=T)}
                if (n7 > 0 )
 {mean7<- mean(sample_base7$cons_exp,na.rm=T)}
```

Calculation variance of consument expenditure for families

```
 if (n1>0)
{var1 <- var(sample_base1$cons_exp,na.rm=T)}
 if (n2>0)
{var2 <- var(sample_base2$cons_exp,na.rm=T)}
 if (n3>0)
{var3 <- var(sample_base3$cons_exp,na.rm=T)}
 if (n4>0)
{var4 <- var(sample_base4$cons_exp,na.rm=T)}
 if (n5>0)
{var5 <- var(sample_base5$cons_exp,na.rm=T)}
 if (n6>0)
{var6 <- var(sample_base6$cons_exp,na.rm=T)}
 if (n7>0)
{var7 <- var(sample_base7$cons_exp,na.rm=T)}
```

#Estimation variance total value

```
if (N1 > 0 & n1 > 0 )
{Dy1=(1-(n1/N1))*(var1/n1) }
if (N2 > 0 & n2 > 0)
 {Dy2=(1-(n2/N2))*(var2/n2) }
if (N3 > 0 & n3>0)
{Dy3=(1-(n3/N3))*(var3/n3) }
if (N4 > 0 & n4>0)
 {Dy4=(1-(n4/N4))*(var4/n4) }
 if (N5 > 0 & n5>0)
{Dy5=(1-(n5/N5))*(var5/n5) }
if (N6 > 0 & n6>0)
 {Dy6=(1-(n6/N6))*(var6/n6) }
  if (N7 > 0 & n7>0)
 {Dy7=(1-(n7/N7))*(var7/n7) }
```

Estimate of standart error

```
Sy1<-sqrt(Dy1)
Sy2<-sqrt(Dy2)
Sy3<-sqrt(Dy3)
Sy4<-sqrt(Dy4)
Sy5<-sqrt(Dy5)
Sy6<-sqrt(Dy6)
Sy7<-sqrt(Dy7)
```

Estimate of total value of sample survey

```
t1<-N1*mean1
t2<-N2*mean2
t3<-N3*mean3
t4<-N4*mean4
t5<-N5*mean5
t6<-N6*mean6
```

Estimation variance for population

```
Dt1<- N1*N1*Dy1
Dt2<- N2*N2*Dy2
Dt3<- N3*N3*Dy3
Dt4<- N4*N4*Dy4
Dt5<- N5*N5*Dy5
Dt6<- N6*N6*Dy6
Dt7<- N7*N7*Dy7
```

#Estimate of standard error

```
St1<-sqrt(Dt1)
St2<-sqrt(Dt2)
St3<-sqrt(Dt3)
St4<-sqrt(Dt4)
St5<-sqrt(Dt5)
St6<-sqrt(Dt6)
```

```
#Total value of household consument expenditure of region
                t<-t1+t2+t3+t4+t5+t6+t7
#Total error of consument expenditure value  of region
                St<-St1+St2+St3+St4+St5+St6+St7
                    j=j+1
                    rt=rt+t
                    rSt=rSt+St
                        }
                    print (rt)
                print
# Relative value of the error in persent
                print (rSt/rt)*100
```

For application in R programming system are used following algorithm:

Average of indicator: $\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$

Estimation variance total value: $D_{\bar{y}} = (1-\frac{n}{N})\frac{S^2}{n}, S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k-\mu)^2$

Estimator of variance: $\hat{D}_{\bar{y}} = (1-\frac{n}{N})\frac{\hat{S}^2}{n}, \hat{S}^2 = \frac{1}{n-1}\sum_{n-1}^{n}(y_k-\bar{y})^2$

:

Estimate of standard error: $\sqrt{\hat{D}_{\bar{y}}} = \hat{S}(\bar{y})$

Estimate of total number of sample survey: $\hat{t} = N\bar{y}$

Estimate of dispersion of population: $\hat{D}\hat{t}$

Estimate of standard error: $\hat{S}(\hat{t})$

Used Reading

1.The Analysis of Household Surveys. A Microeconometric Approach to Development Policy. Reissue Edition with a new preface. Angus Deaton. Winner of the 2015 Nobel Prize in economics. World Bank group.

2. Организации Объединенных Наций. Департамент по экономическим и социальным вапросам. Статистический отдел. Методологические исследования. Серия F. N-98. Составление планов выборки для объследования домашних хозяйств. Нью Йорк, 2010 год.

3.https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

4.https://unstats.un.org/unsd/hhsurveys/pdf/household_surveys.pdf