Sampling Methods Using STATA

Fares Qeadan, Ph.D

Department of Internal Medicine

Division of Epidemiology, Biostatistics, & Preventive Medicine
University of New Mexico Health Sciences Center

September 14, 2015

UNM HEALTH SCIENCES CENTER | Clinical & Translational Science Center

- Probability (Random) Sampling
  - Simple random sampling (SRS)
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling

- Probability (Random) Sampling
  - Simple random sampling (SRS)
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling
- Non-Probability Sampling
  - Convenience sampling
  - Volunteer sampling
  - Judgment (Purposive) sampling
  - Snowball sampling
  - Quota sampling

- Probability (Random) Sampling
    - Simple random sampling (SRS)
    - Systematic sampling
    - Stratified sampling
    - Cluster sampling
    - Multistage sampling
- Non-Probability Sampling
    - Convenience sampling
    - Volunteer sampling
    - Judgment (Purposive) sampling
    - Snowball sampling
    - Quota sampling
- Sampling Bias

- Probability (Random) Sampling
  - Simple random sampling (SRS)
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling
- Non-Probability Sampling
  - Convenience sampling
  - Volunteer sampling
  - Judgment (Purposive) sampling
  - Snowball sampling
  - Quota sampling
- Sampling Bias
- References

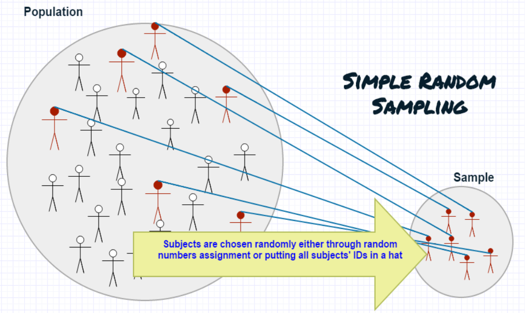
Simple Random Sampling (SRS): it's a sampling method in which each subject of the *sampling frame* has an equal chance of being selected into the sample [1]. SRS is the most popular method of random sampling. There are two types of SRS: with replacement and without replacement. SRS with replacement is less common.

Advantages:

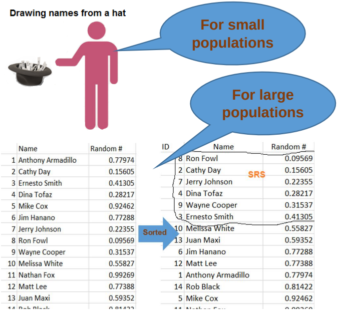
- Easy to use in small populations.
- With an appropriate sample size, SRS provides a highly representative sample of the target population.

Disadvantages:

- Difficult to use in large populations (expensive: time and cost).
- Small segments of the target population may not be present in the sample with sufficient number of subjects.



Subjects are chosen randomly either through random numbers assignment or putting all subjects' IDs in a hat



	Name	Random #
1	Anthony Armadillo	0.77974
2	Cathy Day	0.15605
3	Ernesto Smith	0.41305
4	Dina Tofaz	0.28217
5	Mike Cox	0.92462
6	Jim Hanano	0.77288
7	Jerry Johnson	0.22355
8	Ron Fowl	0.09569
9	Wayne Cooper	0.31537
10	Melissa White	0.55827
11	Nathan Fox	0.99269
12	Matt Lee	0.77388
13	Juan Maxi	0.59352
14	Rob Black	0.81422

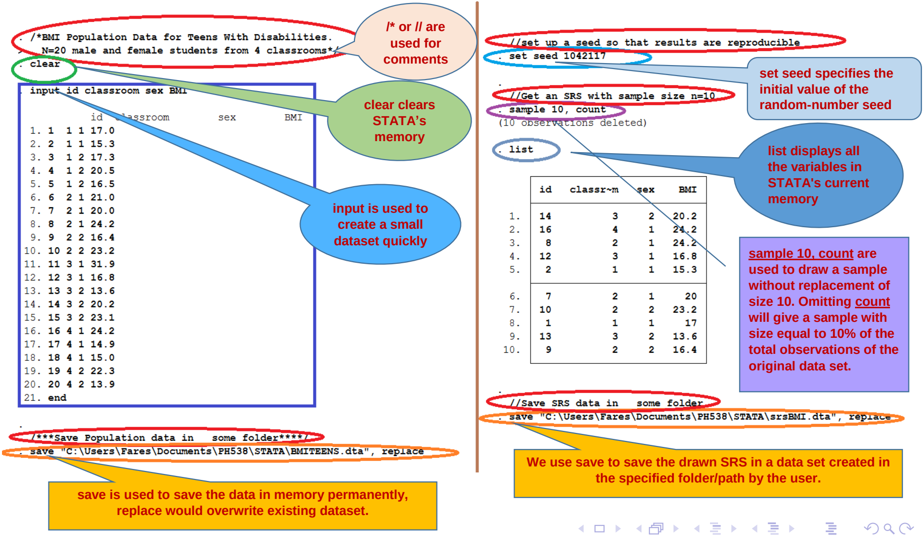
ID	Name	Random #
8	Ron Fowl	0.09569
2	Cathy Day	0.15605
7	Jerry Johnson	0.22355
4	Dina Tofaz	0.28217
9	Wayne Cooper	0.31537
3	Ernesto Smith	0.41305
10	Melissa White	0.55827
13	Juan Maxi	0.59352
6	Jim Hanano	0.77288
12	Matt Lee	0.77388
1	Anthony Armadillo	0.77974
14	Rob Black	0.81422
5	Mike Cox	0.92462
11	Nathan Fox	0.99269

Example: Obtain an SRS of size *n* from the population of all adults over age 50 who have high blood pressure in Albuquerque.

Steps:

1. Enter the sampling frame list into a data set.
2. Assign a random number to each subject.
3. Sort the subjects by the assigned random numbers.
4. Select the first *n* subjects from your sorted list.

How to get an SRS using STATA: In the following example we will obtain an SRS of 10 students from a BMI Population of 20 Teens.



```
. /*BMI Population Data for Teens With Disabilities.
>   N=20 male and female students from 4 classrooms*/
. clear

. input id classroom sex BMI

                 id  classroom         sex        BMI
  1. 1   1 1 17.0
  2. 2   1 1 15.3
  3. 3   1 2 17.3
  4. 4   1 2 20.5
  5. 5   1 2 16.5
  6. 6   2 1 21.0
  7. 7   2 1 20.0
  8. 8   2 1 24.2
  9. 9   2 2 16.4
 10. 10  2 2 23.2
 11. 11  3 1 31.9
 12. 12  3 1 16.8
 13. 13  3 2 13.6
 14. 14  3 2 20.2
 15. 15  3 2 23.1
 16. 16  4 1 24.2
 17. 17  4 1 14.9
 18. 18  4 1 15.0
 19. 19  4 2 22.3
 20. 20  4 2 13.9
 21. end

.
. /***Save Population data in   some folder****/
. save "C:\Users\Fares\Documents\PH538\STATA\BMITEENS.dta", replace
```

/* or // are used for comments

clear clears STATA's memory

input is used to create a small dataset quickly

save is used to save the data in memory permanently, replace would overwrite existing dataset.

```
. //set up a seed so that results are reproducible
. set seed 1042117

.
. //Get an SRS with sample size n=10
. sample 10, count
(10 observations deleted)

. list

     +----------------------------------+
     | id   classr~m   sex    BMI       |
     |----------------------------------|
  1. | 14          3     2   20.2       |
  2. | 16          4     1   24.2       |
  3. |  8          2     1   24.2       |
  4. | 12          3     1   16.8       |
  5. |  2          1     1   15.3       |
     |----------------------------------|
  6. |  7          2     1     20       |
  7. | 10          2     2   23.2       |
  8. |  1          1     1     17       |
  9. | 13          3     2   13.6       |
 10. |  9          2     2   16.4       |
     +----------------------------------+

.
. //Save SRS data in   some folder
. save "C:\Users\Fares\Documents\PH538\STATA\srsBMI.dta", replace
```

set seed specifies the initial value of the random-number seed

list displays all the variables in STATA's current memory

sample 10, count are used to draw a sample without replacement of size 10. Omitting count will give a sample with size equal to 10% of the total observations of the original data set.

We use save to save the drawn SRS in a data set created in the specified folder/path by the user.

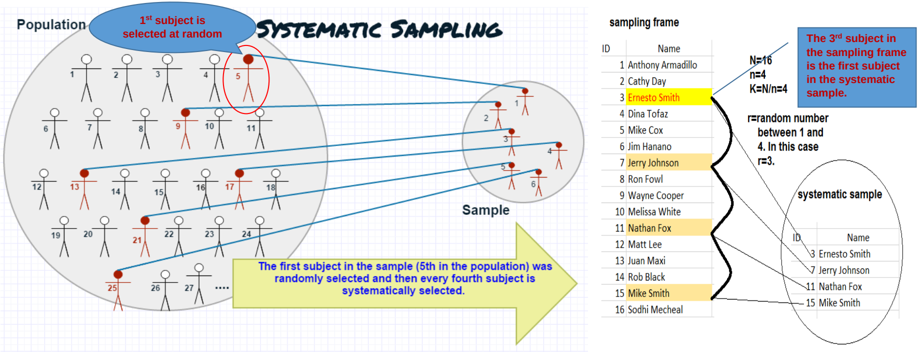
Systematic sampling: it's a sampling method in which subjects are chosen in a systematic way such that one first randomly picks the first subject from the sampling frame and then selects each *kth* subject from the list ($k = N/n$) [1]. If the sampling frame is randomly shuffled, then systematic sampling is equivalent to SRS.

Advantages:

- Fast and easy.
- With an appropriate sample size, it provides a highly representative sample of the target population since, by construction, the sample is uniformly distributed over the sampling frame.

Disadvantages:

- Might lead to bias if the sampling frame is arranged in a specific pattern and the periodicity of the sampling matched the periodicity of that pattern.
- May not capture certain segments of interest from the target population.


Population
1st subject is selected at random
SYSTEMATIC SAMPLING
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
....
Sample
1
2
3
4
5
6
The first subject in the sample (5th in the population) was randomly selected and then every fourth subject is systematically selected.
sampling frame
N=16
n=4
K=N/n=4
The 3rd subject in the sampling frame is the first subject in the systematic sample.
r=random number between 1 and 4. In this case r=3.
systematic sample

sampling frame

ID	Name
1	Anthony Armadillo
2	Cathy Day
3	Ernesto Smith
4	Dina Tofaz
5	Mike Cox
6	Jim Hanano
7	Jerry Johnson
8	Ron Fowl
9	Wayne Cooper
10	Melissa White
11	Nathan Fox
12	Matt Lee
13	Juan Maxi
14	Rob Black
15	Mike Smith
16	Sodhi Mecheal

systematic sample

ID	Name
3	Ernesto Smith
7	Jerry Johnson
11	Nathan Fox
15	Mike Smith

Example: Obtain a systematic sample of size n from the population of all adults over age 50 who have high blood pressure in Albuquerque.

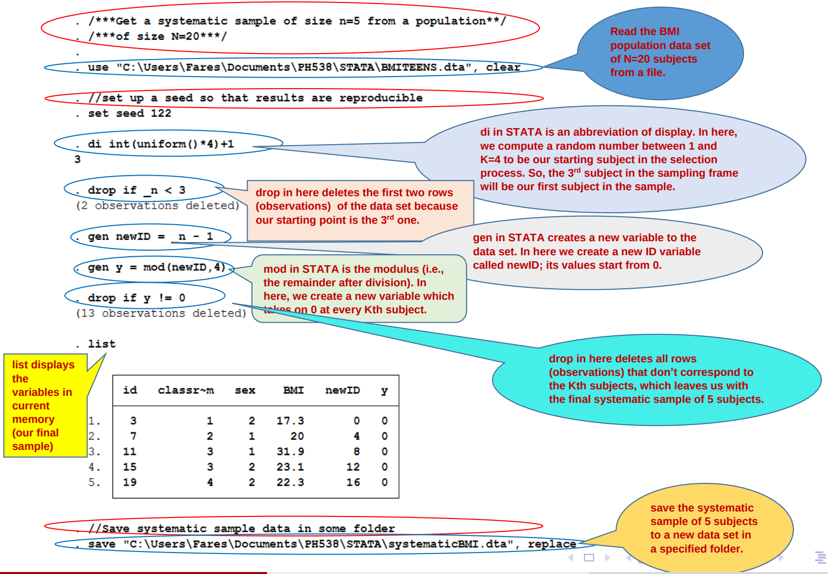
Steps: 1. Enter the sampling frame list into a data set.

2. Calculate the sampling interval $K = N/n$.

3. Generate a random number between 1 and K, say r.

4. Select the r^{th} subject from the sampling frame and then select every K^{th} subject.

How to get a systematic sample using STATA: In the following example we will obtain a systematic sample of 5 students from a BMI Population of 20 Teens.



```
. /***Get a systematic sample of size n=5 from a population**/
. /***of size N=20***/
.
. use "C:\Users\Fares\Documents\PH538\STATA\BMITEENS.dta", clear

. //set up a seed so that results are reproducible
. set seed 122

. di int(uniform()*4)+1
3

. drop if _n < 3
(2 observations deleted)

. gen newID = _n - 1

. gen y = mod(newID,4)

. drop if y != 0
(13 observations deleted)

. list

      id   classr~m   sex    BMI   newID   y

  1.   3          1     2   17.3       0   0
  2.   7          2     1     20       4   0
  3.  11          3     1   31.9       8   0
  4.  15          3     2   23.1      12   0
  5.  19          4     2   22.3      16   0

. //Save systematic sample data in some folder
. save "C:\Users\Fares\Documents\PH538\STATA\systematicBMI.dta", replace
```

Read the BMI population data set of N=20 subjects from a file.

di in STATA is an abbreviation of display. In here, we compute a random number between 1 and K=4 to be our starting subject in the selection process. So, the 3rd subject in the sampling frame will be our first subject in the sample.

drop in here deletes the first two rows (observations) of the data set because our starting point is the 3rd one.

gen in STATA creates a new variable to the data set. In here we create a new ID variable called newID; its values start from 0.

mod in STATA is the modulus (i.e., the remainder after division). In here, we create a new variable which takes on 0 at every Kth subject.

drop in here deletes all rows (observations) that don't correspond to the Kth subjects, which leaves us with the final systematic sample of 5 subjects.

list displays the variables in current memory (our final sample)

save the systematic sample of 5 subjects to a new data set in a specified folder.

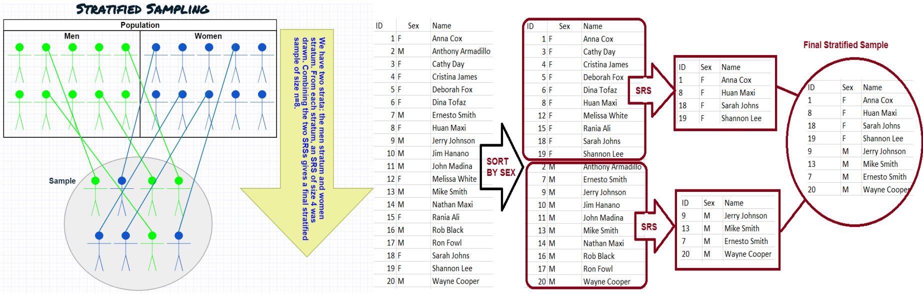
Stratified sampling: it's a sampling method in which a sample is obtained by firstly dividing the population into subpopulations (strata) based on some characteristics and then an SRS is taken from each stratum [1]. Combining the obtained SRSs will give the final stratified sample. Minority subgroups of interest can be ensured by stratification. There are two types of stratified sampling: proportionate and disproportionate. In the proportionate one, we draw a sample from each stratum in proportion to its share in the target population. By this method, each stratum should be internally homogeneous.

Advantages:

- Has the highest precision among other sampling methods.
- The sample is more representative as it allows certain segments of interest, from the target population, to be captured.
- We could use other sampling methods than SRS in each stratum.

Disadvantages:

- Might introduce some complexities at the analysis stage.
- More time consuming and effort than other sampling methods.
- Requires separate sampling frames for each stratum.



We have two strata: the men stratum and women stratum. From each stratum, an SRS of size 4 was drawn. Combining the two SRSs gives a final stratified sample of size n=8.

ID	Sex	Name
1	F	Anna Cox
2	M	Anthony Armadillo
3	F	Cathy Day
4	F	Cristina James
5	F	Deborah Fox
6	F	Dina Tofaz
7	M	Ernesto Smith
8	F	Huan Maxi
9	M	Jerry Johnson
10	M	Jim Hanano
11	M	John Madina
12	F	Melissa White
13	M	Mike Smith
14	M	Nathan Maxi
15	F	Rania Ali
16	M	Rob Black
17	M	Ron Fowl
18	F	Sarah Johns
19	F	Shannon Lee
20	M	Wayne Cooper

ID	Sex	Name
1	F	Anna Cox
3	F	Cathy Day
4	F	Cristina James
5	F	Deborah Fox
6	F	Dina Tofaz
8	F	Huan Maxi
12	F	Melissa White
15	F	Rania Ali
18	F	Sarah Johns
19	F	Shannon Lee
2	M	Anthony Armadillo
7	M	Ernesto Smith
9	M	Jerry Johnson
10	M	Jim Hanano
11	M	John Madina
13	M	Mike Smith
14	M	Nathan Maxi
16	M	Rob Black
17	M	Ron Fowl
20	M	Wayne Cooper

ID	Sex	Name
1	F	Anna Cox
8	F	Huan Maxi
18	F	Sarah Johns
19	F	Shannon Lee

ID	Sex	Name
9	M	Jerry Johnson
13	M	Mike Smith
7	M	Ernesto Smith
20	M	Wayne Cooper

ID	Sex	Name
1	F	Anna Cox
8	F	Huan Maxi
18	F	Sarah Johns
19	F	Shannon Lee
9	M	Jerry Johnson
13	M	Mike Smith
7	M	Ernesto Smith
20	M	Wayne Cooper

Example: Obtain a stratified sample by RACE of size n from the population of all adults over age 50 who have high blood pressure in Albuquerque.

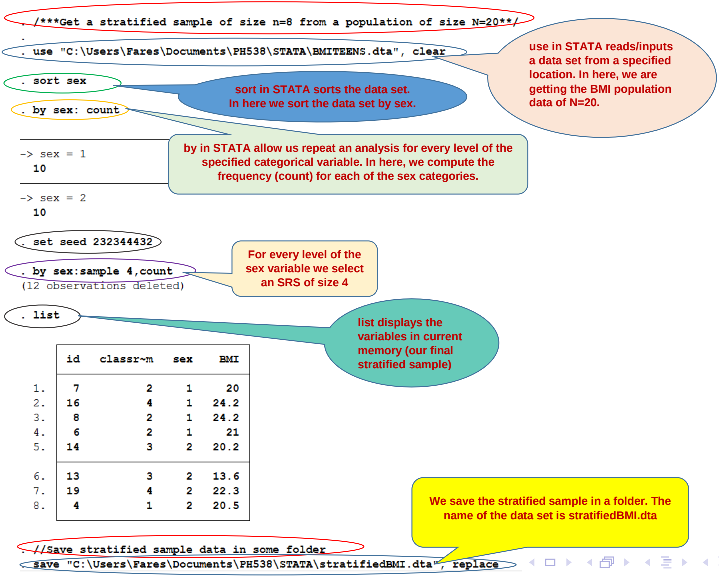
Steps: 1. Enter the sampling frame list into a data set.

2. Sort the sampling frame by Race.

3. Select an SRS from each Race stratum such that the proportion n_i/n reflects the proportion of the i^{th} stratum in the population where n_i is the SRS size obtained from the i^{th} stratum and n is the final stratifies sample size.

4. Combine all SRSs obtained from all strata to make the final stratified sample.

How to get a stratified sample using STATA: In the following example we will obtain a stratified sample by Gender of 8 students from a BMI Population of 20 Teens.



```
. /***Get a stratified sample of size n=8 from a population of size N=20**/
.
. use "C:\Users\Fares\Documents\PH538\STATA\BMITEENS.dta", clear

. sort sex

. by sex: count

-------------------------------------------------------------------
-> sex = 1
  10

-------------------------------------------------------------------
-> sex = 2
  10

. set seed 232344432

. by sex:sample 4,count
(12 observations deleted)

. list

     +------------------------------+
     | id   classr~m   sex     BMI  |
     |------------------------------|
  1. |  7          2     1      20  |
  2. | 16          4     1    24.2  |
  3. |  8          2     1    24.2  |
  4. |  6          2     1      21  |
  5. | 14          3     2    20.2  |
     |------------------------------|
  6. | 13          3     2    13.6  |
  7. | 19          4     2    22.3  |
  8. |  4          1     2    20.5  |
     +------------------------------+

. //Save stratified sample data in some folder
. save "C:\Users\Fares\Documents\PH538\STATA\stratifiedBMI.dta", replace
```

use in STATA reads/inputs a data set from a specified location. In here, we are getting the BMI population data of N=20.

sort in STATA sorts the data set. In here we sort the data set by sex.

by in STATA allow us repeat an analysis for every level of the specified categorical variable. In here, we compute the frequency (count) for each of the sex categories.

For every level of the sex variable we select an SRS of size 4

list displays the variables in current memory (our final stratified sample)

We save the stratified sample in a folder. The name of the data set is stratifiedBMI.dta

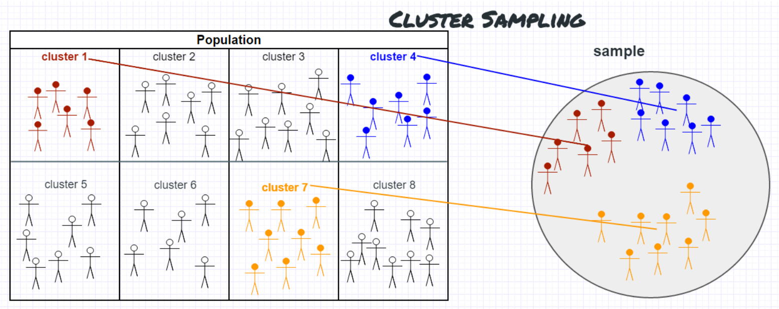
Cluster sampling: it's a sampling method in which the target population is first divided into naturally occurring clusters and then a random sample of clusters is obtained such that all subjects in the randomly selected clusters are included in the sample [1]. Sometimes, we include an SRS from each selected cluster instead of including all subjects which makes the sampling method to be called a two-stage sampling method. By this method, clusters should be internally as heterogeneous as the target population itself.

Advantages:

- Doesn't require a sampling frame.
- Time and cost efficient compared to other sampling methods.
- Cluster samples have larger sample sizes.

Disadvantages:

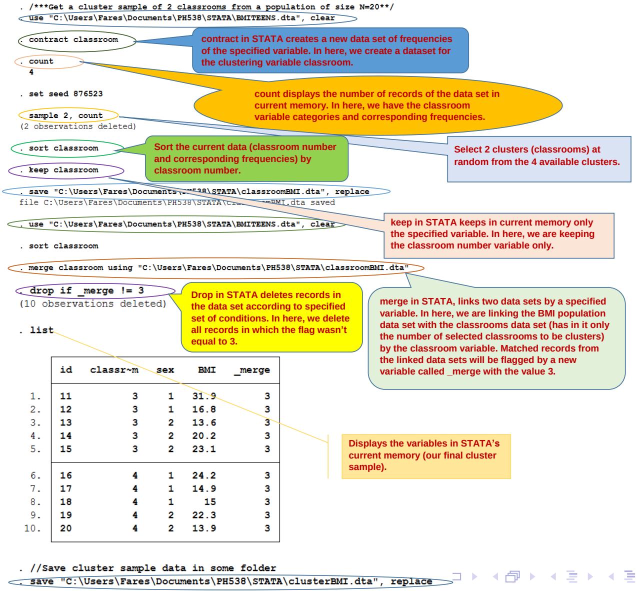
- Produces higher sampling error.
- It's the least representative of the target population among random sampling methods.



Example: Obtain a cluster sample by Geographical Region of size *n* from the population of all adults over age 50 who have high blood pressure in Albuquerque.

Steps: 1. Divide the target population into *k* Geographical Regions (clusters).
2. From the *K* clusters, select at random *h* clusters.
3. For each randomly selected cluster include all subjects (adults over age 50 who have high blood pressure.
4. Combining all selected subjects from the randomly selected clusters makes the final cluster sample.

How to get a cluster sample using STATA: In the following example we will obtain a cluster sample of 2 classrooms of students from a BMI Population of 20 Teens.



```
. /***Get a cluster sample of 2 classrooms from a population of size N=20**/
. use "C:\Users\Fares\Documents\PH538\STATA\BMITEENS.dta", clear

. contract classroom

. count
  4

. set seed 876523

. sample 2, count
(2 observations deleted)

. sort classroom

. keep classroom

. save "C:\Users\Fares\Documents\PH538\STATA\classroomBMI.dta", replace
file C:\Users\Fares\Documents\PH538\STATA\classroomBMI.dta saved

. use "C:\Users\Fares\Documents\PH538\STATA\BMITEENS.dta", clear

. sort classroom

. merge classroom using "C:\Users\Fares\Documents\PH538\STATA\classroomBMI.dta"

. drop if _merge != 3
(10 observations deleted)

. list

      id   classr~m   sex    BMI   _merge

  1.  11          3     1   31.9        3
  2.  12          3     1   16.8        3
  3.  13          3     2   13.6        3
  4.  14          3     2   20.2        3
  5.  15          3     2   23.1        3

  6.  16          4     1   24.2        3
  7.  17          4     1   14.9        3
  8.  18          4     1     15        3
  9.  19          4     2   22.3        3
 10.  20          4     2   13.9        3


. //Save cluster sample data in some folder
. save "C:\Users\Fares\Documents\PH538\STATA\clusterBMI.dta", replace
```

contract in STATA creates a new data set of frequencies of the specified variable. In here, we create a dataset for the clustering variable classroom.

count displays the number of records of the data set in current memory. In here, we have the classroom variable categories and corresponding frequencies.

Select 2 clusters (classrooms) at random from the 4 available clusters.

Sort the current data (classroom number and corresponding frequencies) by classroom number.

keep in STATA keeps in current memory only the specified variable. In here, we are keeping the classroom number variable only.

Drop in STATA deletes records in the data set according to specified set of conditions. In here, we delete all records in which the flag wasn't equal to 3.

merge in STATA, links two data sets by a specified variable. In here, we are linking the BMI population data set with the classrooms data set (has in it only the number of selected classrooms to be clusters) by the classroom variable. Matched records from the linked data sets will be flagged by a new variable called _merge with the value 3.

Displays the variables in STATA's current memory (our final cluster sample).

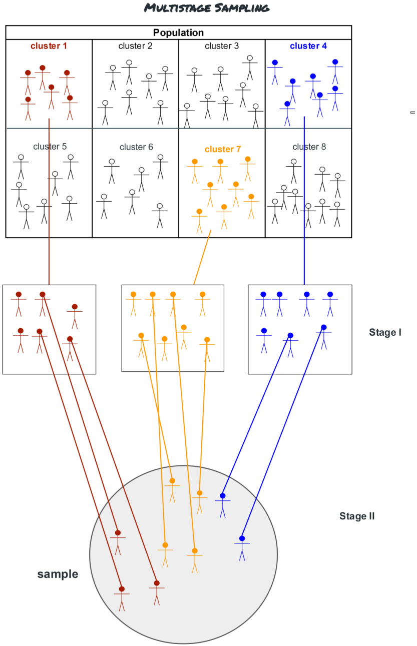
Multistage sampling: It's a sampling method in which we use combinations of two or more sampling methods at least one of which involves randomness [2]. The most common examples for multistage sampling are Stratified random sampling and cluster sampling. For example, in the 2 stage cluster sampling, in Stage 1, we use cluster sampling to choose clusters from a population. Then, in Stage 2, we use simple random sampling to select a subset from each cluster for the final sample.

Advantages:

- Cost and Time effective.
- Sometimes, it does not require a sampling frame.
- Multistage samples have larger sample sizes.

Disadvantages:

- Difficult and complex design.
- Partially subjective.
- Induces lower accuracy due to higher sampling error.


Multistage Sampling
Population
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
cluster 6
cluster 7
cluster 8
Stage I
Stage II
sample

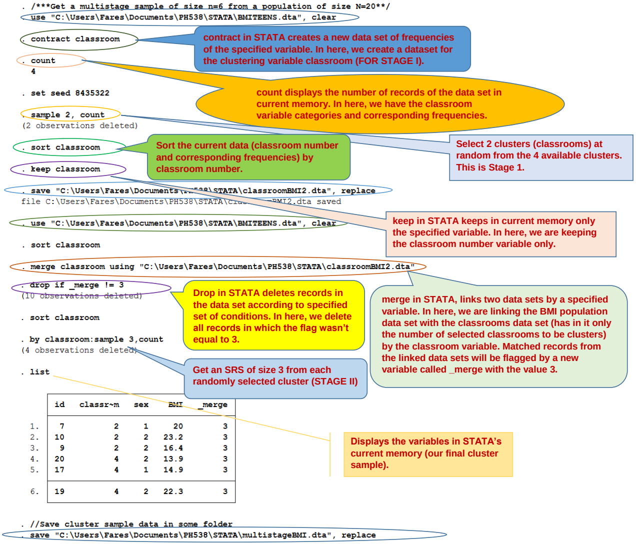
Example 1: Obtain a multistage sample of size *n* from the population of all adults over age 50 who have high blood pressure in Albuquerque.

Steps: 1. Divide the target population into *k* Geographical Regions (clusters).
2. From the *K* clusters, select at random *h* clusters.
3. From each randomly selected cluster select an SRS.
4. Combining all selected SRSs makes the final multistage sample.

Example 2: Obtain a multistage sample of size *n* from the population of all adults over age 50 who have high blood pressure in Albuquerque.

Steps: 1. Enter the sampling frame list (if available) into a data set.
2. Sort the sampling frame by Race.
3. Select an SRS from each Race stratum such that the proportion n_i/n reflects the proportion of the i^{th} stratum in the population where n_i is the SRS size obtained from the i^{th} stratum and n is the final stratifies sample size.
4. Combine all SRSs obtained to make the final multistage sample.

How to get a multistage sample using STATA: In the following example we will obtain a multistage sample of size 6 from a BMI Population of 20 Teens.



```
. /***Get a multistage sample of size n=6 from a population of size N=20**/
. use "C:\Users\Fares\Documents\PH538\STATA\BMITEENS.dta", clear

. contract classroom

. count
  4

. set seed 8435322

. sample 2, count
(2 observations deleted)

. sort classroom

. keep classroom

. save "C:\Users\Fares\Documents\PH538\STATA\classroomBMI2.dta", replace
file C:\Users\Fares\Documents\PH538\STATA\classroomBMI2.dta saved

. use "C:\Users\Fares\Documents\PH538\STATA\BMITEENS.dta", clear

. sort classroom

. merge classroom using "C:\Users\Fares\Documents\PH538\STATA\classroomBMI2.dta"

. drop if _merge != 3
(10 observations deleted)

. sort classroom

. by classroom:sample 3,count
(4 observations deleted)

. list

     +--------------------------------------+
     | id   classr~m   sex    BMI   _merge |
     |--------------------------------------|
  1. |  7          2     1     20        3 |
  2. | 10          2     2   23.2        3 |
  3. |  9          2     2   16.4        3 |
  4. | 20          4     2   13.9        3 |
  5. | 17          4     1   14.9        3 |
     |--------------------------------------|
  6. | 19          4     2   22.3        3 |
     +--------------------------------------+

. //Save cluster sample data in some folder
. save "C:\Users\Fares\Documents\PH538\STATA\multistageBMI.dta", replace
```

contract in STATA creates a new data set of frequencies of the specified variable. In here, we create a dataset for the clustering variable classroom (FOR STAGE I).

count displays the number of records of the data set in current memory. In here, we have the classroom variable categories and corresponding frequencies.

Select 2 clusters (classrooms) at random from the 4 available clusters. This is Stage 1.

Sort the current data (classroom number and corresponding frequencies) by classroom number.

keep in STATA keeps in current memory only the specified variable. In here, we are keeping the classroom number variable only.

Drop in STATA deletes records in the data set according to specified set of conditions. In here, we delete all records in which the flag wasn't equal to 3.

merge in STATA, links two data sets by a specified variable. In here, we are linking the BMI population data set with the classrooms data set (has in it only the number of selected classrooms to be clusters) by the classroom variable. Matched records from the linked data sets will be flagged by a new variable called _merge with the value 3.

Get an SRS of size 3 from each randomly selected cluster (STAGE II)

Displays the variables in STATA's current memory (our final cluster sample).

Convenience sampling: it's a non-probability sampling method in which subjects are conveniently available to the researcher [3]. This is the most popular method of non-probability sampling and sometimes is called accidental sampling.

Advantages:

- Cheap and simple; requires no planning.
- Helpful for pilot studies and hypotheses generation.

Disadvantages:

- Unrepresentative of the target population.
- Suffers from selection bias.

Remark: For other non-probability sampling methods please revisit [4].

Sampling Bias: Sampling bias refers to over-representation or under-representation of some subgroups of the target population. There are two types of sampling bias including random errors and systematic errors [5].

Random Error: error is reduced with increased sample size. It's due to the sample size. Error is evenly distributed across the sampling frame.

Systematic Error (bias): error is not reduced with increased sample size. It's due the design; mainly non-randomness.

References

- [1]. Lohr, Sharon (2009). Sampling: design and analysis. Cengage Learning.
- [2]. Foreman, E. K. (1991). Survey sampling principles. CRC Press.
- [3]. Trevino, J. J. (2012). Addiction Research Methods edited by Peter G. Miller, John Strang, Peter M. Miller.
- [4]. DePoy, E., and Gitlin, L. N. (2015). Introduction to research: Understanding and applying multiple strategies. Elsevier Health Sciences.
- [5]. Norell, S. E. (1995). Workbook of epidemiology. Oxford University Press.

How to cite this work:

This work was funded by the NIH grants (1U54GM104944-01A1) through the National Institute of General Medical Sciences (NIGMS) under the Institutional Development Award (IDeA) program and the UNM Clinical & Translational Science Center (CTSC) grant (UL1TR001449). Thus, to cite this work please use:

Fares Qeadan (2015). Sampling Methods Using STATA. A short course in biostatistics for the Mountain West Clinical Translational Research Infrastructure Network (grant 1U54GM104944) and UNM Clinical & Translational Science Center (CTSC) (grant UL1TR001449). University of New Mexico Health Sciences Center. Albuquerque, New Mexico.

Thank you.

For questions, Email: FQeadan@salud.unm.edu

For STATA:

Do file: http://www.mathalpha.com/SAMPLING/sampling.do