



Forwarding Armenian Statistics Through Twinning

AM09/ENP-PCA/TP/04

MISSION REPORT

on

Quality management



A4.3 Sampling methodology

Mission carried out by

Dr. Kai Lorentz, Destatis

Armenia, 10-14 September 2012

Final version

		
National Statistical Service Republic of Armenia		The German Federal Statistical Office

Dr. Kai Lorentz

Group of Mathematical-Statistical Methods

Destatis – Statistisches Bundesamt

Gustav-Stresemann-Ring 11

DE-65189 Wiesbaden

Germany

kai.lorentz@destatis.de

www.destatis.de

List of Abbreviations

MS Member State

NSSRA National Statistical Service of the Republic of Armenia

QWG Working Group on Quality

Table of contents

1	General comments	4
2	Actions and results	4
3	Conclusions and recommendations	4
3.1	Presentation on sample survey errors	4
3.2	Demonstration of the usage of the R software	5
3.3	Consultation on sample survey issues	7
3.4	Discussion about the need to have a Sampling Expert Group within the Statistical Office of Armenia	8
4	Actions before next mission	8

Annexes

Annex 1	Terms of Reference	9
Annex 2	Meeting programme	10
Annex 3	Persons met	11
Annex 4	Power Point presentation on sample survey errors	12

1. General comments

This report was prepared as the result of a mission to the National Statistical Service of the Republic of Armenia. It was the author's first visit to the National Statistical Service of the Republic of Armenia.

The MS Expert would like to express warm thanks to all officials and individuals met for the kind support and valuable information, which was received during the stay in Armenia, and which highly facilitated the work.

The views and observations stated in this report are those of the MS Expert, and do not necessarily correspond to the views of Destatis.

2. Actions and results

During the mission the expert was involved in the following actions:

- Outline the importance of Standard Error Estimations of sample survey results for Quality Management of Statistics.
- Explaining the concepts of Standard Error Estimations to NSSRA staff members.
- Demonstration of the usage of the R-software to carry out Standard Error Estimations for the example of the Labor Survey.
- Consultation in Sample Survey Issues: Transportation Statistics.
- Discussions about organizational structure of Armenian Statistical Office, especially about the need to have a Sampling Expert Group.

3. Conclusions and recommendations

3.1 Presentation on Sample Survey Errors

A presentation on types of errors related to sample surveys was given to NSSRA staff members. Its obvious importance in the context of quality management was underlined. The slides of the presentation can be found in Annex A4.3.4. The presentation was translated into Armenian, and was made available to all NSSRA staff members.

It is highly recommended to carry out Standard Error Calculations regularly.

3.2 Demonstration of the usage of the R-software:

The statistical software package “R” is becoming more and more popular for statistical computing, not only in academic institutions and universities, but also in national statistical offices.

In the context of analysis of sample survey data, e.g. the calculation (estimation) of Standard Errors, the package “survey” was applied during the MS Expert’s visit to NSSRA and its usage was explained to NSSRA's staff members. Below, the most important steps and commands of the analysis undertaken, so that repetition of the calculations carried out is possible. For a deeper learning and understanding of the “survey” package, it is recommended to consult the book: “*Complex Surveys, a guide to Analysis using R*” by Thomas Lumley, *Wiley series in survey methodology*.

In the sequel, we describe the undertaken steps (as an example we take here the Labor Survey and some of its variables, the file was named “Labour survey - final stage and unweighted survey results.xlsx”):

1. Preparation of the survey data:

The survey data file has to contain all the relevant variable values of study interest, in our case: variables named as h_x_y (e.g. h_310_1).

It has to contain identifier variables for the statistical units: in our case of a one-stage sampling design of establishments.

It also has to contain variables about the stratification: in our case it was named Strata.

Moreover, the input data set has to include the weights-variable, the variable containing the adjusted weights of the units (variable Weights).

Last, it has to contain a variable indicating the sub-populations of study interest, in our case the Administrative Regions, variable ID_MARZ.

The data has to be sorted according to the stratification variable.

The prepared input data file (input of the following analysis performed) has to be stored as a “.csv”-text file, here “Labour survey - final stage and unweighted survey results.csv”.

2. Starting the application “R”.

3. Choose the working directory for the project:

Menu item “File → Change dir...” and browse to your folder of choice.

4. Download the R-package “survey” from

<http://www.r-project.org/> → CRAN → <http://ftp5.gwdg.de/pub/misc/cran/> → Packages → http://ftp5.gwdg.de/pub/misc/cran/web/packages/available_packages_by_name.html → S → survey (http://ftp5.gwdg.de/pub/misc/cran/bin/windows/contrib/r-release/survey_3.28-2.zip).

It is recommended to download also the manual “survey.pdf”

(<http://ftp5.gwdg.de/pub/misc/cran/web/packages/survey/survey.pdf>).

5. Install the R-package “survey”:
Menu item “*Packages → Install package(s) from local zip files ...*” and browse to previously downloaded “survey”-package (downloaded as zip file).
6. Make the additional functions, defined in the “survey”-package known to the R-system: this is done by the following command:

```
> library (survey)
```

 (the “>” is the R-prompt).
7. Read in the input data file (“.csv”-text file created in step 1 and copied to the working directory of the project, here it is called “Labour survey - final stage and unweighted survey results.csv”) and assign it to an R-object (here called `exampledata`):

```
> exampledata <- read.csv ("Labour survey - final stage and unweighted survey results.csv")
```

 If you want to have a brief look on the dataframe “exampledata”, you can type the command

```
> summary (exampledata)
```
8. Create a survey design object for the sample data in question and assign it to an R-object (here called `exampledesign`):

```
> exampledesign <- svydesign (id=~1, strat=~Strata, weights=~Weights, data=exampledata)
```
9. In order to handle strata with only one sampling unit, one has to set the following option:

```
> survey.lonely.psu = "remove"
```

 this choice of option assumes that strata with only one sampling unit are strata with weight equal to 1.
10. Performing the calculation (estimation) of standard errors:
 First, we do this for the totals of the variables of interest at Contry level: this is done by

```
> svytotal (~h_310_1, design=exampledesign)
```


 Second, we carry this out for the totals at the level of the 11 Administrative Regions:>

```
svyby (~h_310_1, by= ~ID_MARZ, svytotal, design=exampledesign)
```


 Next, we calculate the standard error for the ratio of the totals of h_310_1 over the total of h_320_1 at the country level:

```
> svyratio (~h_310_1, ~h_320_1, design=exampledesign)
```

11. Last, we calculate the standard errors for these quotients at the level of the 11 provinces:

```
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 1))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 2))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 3))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 4))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 5))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 6))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 7))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 8))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 9))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 10))
> svyratio (~h_310_1, ~h_320_1,
design=subset(exampledesign, ID_MARZ == 11))
```

Now, also relative standard errors may be computed from the above calculated (absolute) standard errors.

12. At the end, results and R-session of the project may be saved to repeat it next time:
Menu item “*File → Save Workspace...*” and browse to your folder of choice
Menu item “*File → Save History...*” and browse to your folder of choice
Menu item “*File → Save to File...*” and browse to your folder of choice.

3.3 Consultation In Sample Survey Issues

Sampling related issues were discussed for Transport Statistics. As the frame population at the moment cannot be identified out from the Business Register, a sampling approach was discussed in order to estimate the total number of establishments related to transportation and the total number of trucks.

It was pointed out, that this information at the end could be used to re-weight the results of the current survey.

Also, sampling plan issues were discussed: stratification and sample size allocation.

3.4 Discussion about the need to have a Sampling Expert Group within the Statistical Office of Armenia

Nowadays it is crucial for National Statistical Offices, to have a very good knowledge of sampling methodology. This is because financial and personal resources of the NSOs are limited and statistics production has to be effective and punctual. Also response burden of enterprises and households is a general topic with which NSOs are confronted.

Sampling techniques – if properly applied – provide a solution to this challenge. They help in cost saving, timely statistics production and to lower the response burden.

Obviously, one has to deal with sampling errors and to care about the quality of statistical output. To discuss about this issue, was one aim of this mission.

Survey planning, especially from the mathematical-statistical point of view, is an important task in every statistical office nowadays. It is therefore important to have specialized and educated staff to perform these tasks at a level, which is state of the art compared to the currently used methods by international statistical offices.

It is not possible, even for larger statistical offices, to have such specialists in every subject matter area. Also, as statistical methodology is developing rather quickly, one can not at a time learn the newest methods from literature and work on subject matter tasks.

It is therefore recommended to have a group (at least 3-5 persons) of mathematical statistical specialists in IT/Research Department, which work independently from the subject matter areas and provide consultation to the subject matter divisions of NSO in the development of sampling plans for surveys, the estimation procedures, the statistical disclosure control and the performing of seasonal adjustment.

Such a group serves as a “knowledge centre” for mathematical-statistical purposes to the subject matter divisions of NSO.

4. Actions before next mission

The next mission planned within the Quality Management component is the review mission, A.6. However, there is a chance that an additional mission to Armenia on sampling issues may be implemented – either before A.6 or in connection hereto (to be decided).

Before the next mission it is expected that NSSRA continues the work with the R software in relation to estimating sampling errors, both with the example already initiated (labour survey) plus another sample.

This will help the staff of NSSRA getting acquainted with the methodology.

The results from this work and possible questions to the MS Experts should be provided as a short report to the project’s RTA three weeks before the next mission in order to be able to get final comments and recommendations from the MS Experts.

Annex 1. Terms of Reference

Activity A4.3 Training of staff (3)

1. Purpose of activity

During activity A4.3, staff of the Methodology Division of NSSRA will first of all be trained on the important issue of sampling error measurement, to be included in the quality declaration as one important indicator of quality. In addition hereto, the organization of the future co-operation on sampling issues between the Methodological Division and subject matter divisions of NSSRA will be laid out during the activity.

2. Expected output of the activity

The expected output of the activity is:

- A general introduction to the concept of sampling errors to the members of the NSSRA working group on quality management and key members of the subject matter divisions;
- The sampling error estimated for one statistics in co-operation with the methodology unit of NSSRA and the relevant subject matter division;
- The methodology described in a general and systematic way and translated into Armenian;
- Agreement on a draft time plan for having a sampling error estimate for all statistical products of NSSRA;
- Agreement on proposal on the organisation of the future methodological work within NSSRA on sampling error estimation;

3. Project Participants

Mr Stepan Mnatsakanyan, President of NSSRA (*BC Component Leader*)

Ms Anahit Safyan, Head of International Statistical Cooperation Division

Ms Lilit Petrosyan, Head of Methodological Division

Mr Layert Haratyunyan, Head of Business Register Division

Mr Kai Lorentz, Head of Section, Methodological Division, Destatis (Federal Statistical Bureau of Germany)

Annex 2. Program

Time	Place	Event	Purpose / detail
Monday, 10 Sept. Morning	Congress Hotel	Meeting with RTA	To discuss and plan in detail the programme of the week
Afternoon	NSSRA	Meeting with BC Component Leader and Experts	The relevance of indicators of sampling errors to quality management. BC presentation of alternative statistics to use for case study, and decision of what statistics to work with during the week.
Tuesday, 11 Sept. Morning	NSSRA	Workshop with subject matter division of NSSRA	MS Expert presentation on the concept of sampling error as distinguished from other types of errors. Discussion.
Afternoon	NSSRA	Meeting with BC Experts on Methodology	Case study. Estimating the sampling error for a representative statistical product of NSSRA
Wednesday, 12 Sept.	NSSRA	Meeting with BC Experts on Methodology	Case study, continued
Thursday, 13 Sept. Morning	NSSRA	Meeting with BC Component Leader	Discussion of organisational aspects related to future work and procedures with respect to sampling error measurement
Afternoon	NSSRA	Meeting with BC Experts on Methodology	Drafting of time plan for sampling error estimates for all the statistical products of NSSRA
Friday, 14 Sept. Morning	NSSRA	Workshop with subject matter divisions	MS and BC Expert joint presentation on preliminary results from case study
	NSSRA	Ad hoc meetings	Drafting of report and preparation of presentation on conclusions, recommendations and actions to be taken before next mission.
	NSSRA	Debriefing with BC Project Leader	Conclusions, decisions and recommendations. Consequences for the next mission and implied work programme for BC Experts

Annex 3. Persons met

Stepan Mnatsakanyan, President, NSSRA

Lilit Petrosyan, Member of State Council on Statistics

Anahit Safyan, Head of International Statistics Division, NSSRA

Laert Harutyunyan, Head of Methodology Working Group, NSSRA

Suren Khachatryan, IT Division specialist, NSSRA

Lusine Kalantaryan, Labor Statistics division head

Vardan Arevshatyan, Agriculture statistics division specialist

Armine Shaboyan, Methodological division, first class specialist

Inga Baroyan, Trade and other services statistics division head

Haykush Titizyan, Finance statistics division head

Diana Martirosova, Household survey division head

A Tokhatyan, Population Census and demography division specialist

Georgi Torosyan, Transport and communication statistics division head

Arsen Avagyan, Agriculture statistics division head

Ashot Ananyan, Industry statistics division head

Gurgen Martirosyan, Price statistics and international comparisons division head

Annex 4. Power Point presentation on “Sampling errors”

Errors in survey samples

There are several reasons why statistical estimates do not coincide with the „true“ value. We distinguish between two main error components:

- (1) errors related to sampling
- (2) non-sampling errors

In census surveys, we have only errors of the second kind (2), in sample surveys additionally errors of the first category (1).

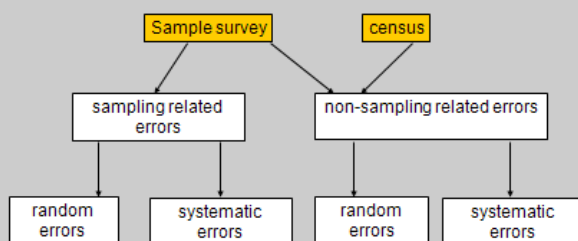
The errors from both categories may be separated into

random errors

and

systematic errors (bias)

Kinds of errors:



Random errors

If survey (drawing of sample etc.) would be repeated several times, estimates would differ according to influence of randomness

⇒ random errors

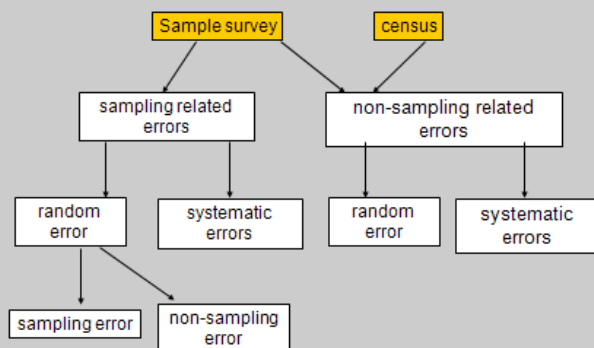
- *Magnitude and direction of randomness error are not known from the beginning*

⇒ random errors can not be compensated

In random sampling the random errors occur, because only part of information from the frame population is known and this part is by chance

⇒ sampling error

Kinds of errors:



- Enlarging sample size reduces the sampling error.
- In random sampling the magnitude of the sampling error can be estimated out of the sample.

result: standard error

Example:

Portion of jobless people is estimated through a simple random sample. Estimated value of this portion: $\hat{p} = 10\%$
It is valid that:

With a probability of 68% the true value of this portion in the population is in the interval

[0,5 %; 19,5 %]	if	10 persons are in the sample
[7 %; 13 %]	if	100 persons are in the sample
[8,7 %; 11,3 %]	if	500 persons are in the sample
[9,7 %; 10,3 %]	if	10000 persons are in the sample

• **Convention:**

We speak of *precision* of the estimated result, when we refer to the error caused by the effect of randomness.

Systematic Errors (bias)

The bias has its origin not in randomness effects, but in systematic errors of data collection and data processing.

If a survey would be repeated several times

⇒ estimated results would differ systematically (in magnitude and direction) from the true value.

Example: Systematic measurement errors, survey non-response.

- Systematic errors can not be estimated out of the survey data. In practice, control surveys are needed.
- Systematic errors may be reduced through careful planning and processing of the survey.
- Enlarging the sample size does not reduce the bias.
- In census surveys the bias is potentially higher than in sample surveys.

Resulting error

is defined as

Resulting error

RMSE =

$$((\text{random error})^2 + (\text{bias})^2)^{1/2}$$

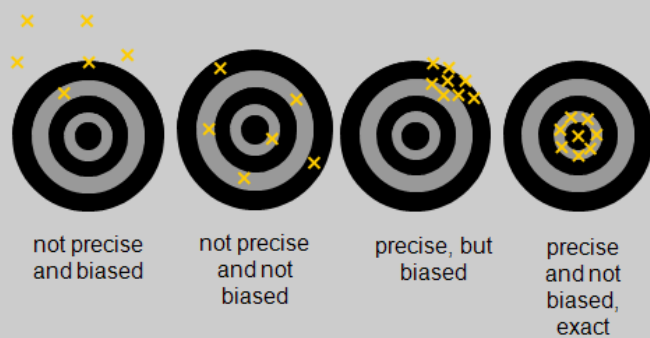
- In sample surveys one should achieve a balanced relationship between the error types.

Example:

Random error	Bias	Resulting error
%	%	%
1,0	10,0	10,0
2,0	10,0	10,2
4,0	10,0	10,8
8,0	10,0	12,8
10,0	10,0	14,1

© Statistisches Bundesamt, C 1

Quality of results:



© Statistisches Bundesamt, C 1

Aim of survey planner:

Under given constraints (monetary and personal resources, time etc.)

avoid *bias*

and

minimize random *errors*!

© Statistisches Bundesamt, C 1

Sources of error

Sampling related errors

bias

- frame imperfections
- sampling method
- sampling technique
- estimation procedures

random errors

non sampling related errors

In preparation phase In survey conduction In survey processing

- | | | |
|---|----------------|-----------------------|
| - planning of sample | - responders | - coding |
| - preparatory work | - interviewers | - data copying errors |
| - inadequate personal and financial support | - form filling | |

- | | | |
|-----------------------------------|-------------------------------------|-------------------------------|
| - organisation of survey | - collection of questionnaire forms | - plausibility checks |
| - questionnaire | | - calculation of results |
| - printing of questionnaire forms | | - production of result tables |
| - training of interviewers | | - interpretation of results |

Non-sampling related errors

Data collection errors

Frame imperfections

Units

- are not within the frame
- are more than once listed in the frame
- do not belong to the population in question

Non-response

Units are

- not found
- not at home
- refuse to answer

Questionnaire form

- is lost
- is not accepted by data processing unit

When is a sample representative?

Usual definition in the literature: a sample is representative, if its distribution of relevant variables correspond to the distribution in the population (frame).

Disadvantage of this definition:

- Not very precise.
- Not suitable for quality measurement purposes.

Alternative definition of representativeness in survey samples:

A sample survey is considered as representative, if standard errors of estimated results are at an acceptable level.

Dissimination praxis at DESTATIS:

- Estimates with a relative standard error higher than 15% are not published (slash).
- Estimates with a relative standard error between 10 and 15% are published only with a footnote, indicating the poor quality of the result.