*National Statistical Service of Greece*


# Sample of Greek population census data

**Stefanos G. Giakoumatos[1] and Ioannis Nikolaidis[2]**

## 1. INTRODUCTION

Population censuses, conducted every 10 years, aim at a complete enumeration of those who live in Greece at a given moment, and also intended to yield picture of the inhabitants' demographic, economic and other characteristics.

In detail, the censuses of population and housing in Greece have three main targets (Redfern 1987):

a. To provide official figures of population
b. The census serves the constitutional function of counting the people that are registered as electors in each Nomos (prefecture). These figures determine the each area's representation in the National Parliament
c. The census provides statistics of the demographics, social and economic characteristics of the population and statistics of housing at national and local levels.

The government body that organizes and conducts the censuses in Greece is the **National Statistical Service of Greece (NSSG)**. The NSSG is directly responsible to the Minister of Economy and Finance, but it is an "independent office" within this ministry. The NSSG budget and personnel cover functions carried out by the central division of the Ministry, by the statistical services located in other ministries and by the statistical offices in each of the 50 Nomos (equivalent to county or prefecture).

[1] Dr. S. Giakoumatos is officer of the Methodology, Analysis and Research sector of the National Statistical Service of Greece; email: giakoumatos@statistics.gr
[2] Mr. I. Nikolaidis is head of the Methodology, Analysis and Research sector of the National Statistical Service of Greece; email: giannikol@statistics.gr

## 2. THE CENSUS PROCEDURE

In this paragraph, a brief account will be given of the procedures followed in the general census.

As a base for the planning of the fieldwork of censuses, the Greek Air force conducts an aerial photographic survey of the Greece. Using these results, maps are created and on these maps the whole country was divided as follows:

a. Each Nomos (prefecture) was divided into enumeration sectors
b. Each enumeration sector was further subdivided into **enumeration districts**

The enumeration district consisted of an area containing about 40-50 households and was under the charge of one enumerator.

On the day of the Census the enumerator, using a sketch of his district, visited the dwellings and all other places of abode in his district and enumerated all the households living in them.

For each household, the enumerator was required to fill in a questionnaire **(Form P1).** The persons to be enumerated in the household were the following:

a. Head of the household
b. Other members of the household including visitors who had spent the night before the census day with the household
c. Members of the household who did not spend the night of census day with the household due to night work, and
d. Persons coming from a journey and arriving at the household on the census day provided that they had not enumerated elsewhere.

Each enumerator prepares a **summary list (form K1)** of the households enumerated by him in his district, The list gave, for each household, the address and the name of the head among with some other particulars.

Institutional households (such as hotels, hospitals, asylums, military barracks etc) consisted the so-called "special enumeration districts". A separate **questionnaire - form P2** - was used for the institutional households in general, irrespective of size.

The small collective households, which were not considered as special enumeration districts were inserted in the list K1 of the appropriate ordinary enumeration districts.

The procedure of the census may now summarized as follows:

i. On the day of the Census the enumerator had to visit one by one all the dwellings included in his district.
ii. For each dwelling he had to enumerate all the persons, who spent the previous night in the dwelling

After the census –three/four days later- a coverage study takes place. The coverage study of the population census is designed to serve as a check on both i and ii points on a sample basis.

## 3. COVERAGE STUDY

### 3.1 General

Understanding of wide scope of a census, it can scarcely reach ideal accuracy and complete freedom of errors (from both administrative and technical points of view). The degree of success of the whole effort would finally depend on the qualifications of those who actually carry out the census that is on the ability, training, judgment and "honesty" of the interviewers.

The possible errors occurring in a general population census, can be classified in the following two general categories:

a. Errors concerning the coverage of the census
b. Errors concerning the content of the census questionnaires.

For the aforementioned reasons, a coverage study follows the Census, in order to estimate the errors that took place during the Census.

### 3.2 The sample design

The dwellings within an enumeration district can be considered as belonging to one of the following two categories

a. Those which were missed at the census and
b. Those which were actually visited by the enumerator

In the first case the coverage error is due to the failure on the part of the enumerator to list the dwelling whereas, in the second case, the error is due to the fact that the enumerator could not include those and only those persons who should have been enumerated.

The following sampling procedures were applied for checking on (a) and (b).

1. We select a sample of areas (enumeration districts). The dwellings within the selected area are re-surveyed to see if any were missed at the census. For those missed, the interviewers fills in a questionnaire. In this way, by means of area sampling, the percentage of dwellings, which were missed on the census, is estimated.
2. The sample of dwellings, which were included in the census, will be selected from the summary list K1. A systematic sample of households is selected from this list and the address and the full name of the household head is copied out to serve identification particulars of the dwellings in the sample. The interviewer visiting the selected dwelling has to collect information for the household in the sample and for any other household staying at the selected dwelling.

**3.3 Stratification**

The whole country was divided in major strata on the basis of population of previous census, as follows:

- 1st stratum: Agglomerations and Municipalities with 40.000 inhabitants or more
- 2nd stratum: Municipalities with 10.000-39.999 inhabitants
- 3rd stratum: Municipalities with 5.000-9.999 inhabitants
- 4th stratum: Municipalities with 2.000-4.999 inhabitants
- 5th stratum: Rest of Greece

The above major strata were divided into 94 equally sized sub-strata.

**3.4 Sample selection**

The sample procedure is as follows:

- 1st stratum: From each substratum of the 1st stratum, a sample of 2 primary sampling units (enumeration sectors) was drawn with replacement and with probabilities proportionate to the size of the units (number of inhabitants according to the previous census).
- 2nd, 3rd ,4th and 5th stratum: From each sub-stratum, 2 primary sample units (municipalities and Communes) were selected with replacement and with probabilities proportionate to the size of the units. In each selected primary unit one secondary sampling unit (enumeration sector) is selected with probability proportionate to the size of unit.

In all the above strata, in each enumeration sector, a sample of three enumeration districts is selected randomly with equal probabilities. From the three enumeration districts, one was used for the detection of the missing households (area sample investigation). In the rest two enumeration-districts, a systematic sample of dwellings was selected (one out of five) from the summary list of K1, for checking the quality of collected data through the Census (list sample investigation).

**3.5 Summary description of the survey procedures**

The survey interviewers had been assigned to areas other than those in which they had any responsibility for the conduct of the general census.

The sequence of the activities of the interviewers can be described as follows:

*A) Area sampling investigation*

The sample of enumeration districts of the sample was identified on the ground with the help of the map provided. The interviewer, after identifying the limits of the enumeration districts of the sample, visited one after the other all the dwellings and other places of the abode and wrote down in a special form (DE1) the following identifications data:

a. Address (street and number locality)
b. Characterization of dwelling (normal, institutional etc)
c. Name and surname of the head of the household
d. Name and surname of the last occupant in case the dwelling was found vacant

A check was then carried out with the census list (K1) for the enumeration district to find out whether the dwellings contained in the list DE1 had been included

in the list K1. For the dwellings, which were missed during the population census, the survey interviewer was to fill in the questionnaire of the survey, after a personal interview with its inmates.

*B) List sample investigation*

The interviewer selected from the list K1 of the two enumeration districts in the list sample (1 out of 5 dwellings) applying systematic sample scheme. Then the interviewer copied down the special form DK1 the identification particulars of the selected households. These identification data (street and number, full name etc) helped the survey interviewer to locate the dwelling intended for the investigation and fill in the questionnaire by means of an interview.

**3.6 Description of the estimation techniques**

We start with the simplest of the cases to be examined, that is the estimation of the coverage error based on the method which consists of calculating on the basis of the census definition and for each dwelling of the list sample or for each part of the area sample separately.

A difference on the following sense:

- As proved by the survey, this dwelling included x′ persons in the Census day, which ought to be enumerated in accordance with the census instructions.
- The census interviewer enumerated x persons
- A comparison of the figures would be reveal a difference or a net error y= x′-x.

This method is a simple and easy one, because it requires nothing more than a comparison of the survey questionnaire with the corresponding census questionnaire. A suitable mathematical formula would make possible to estimate the total coverage error out of the errors of the estimation y of the sample units.

It is recalled that the sample design was a multi-staged one (3-stages in the 1st stratum and 4-stages in the rest strata) and stratified on the first stage. The estimation process will be described for the first stratum. For the rest strata, although the 4-stage sampling is applied, the process is approximately the same.

The symbols used are as follows:

- ✓ h: The class of the sub-strata
- ✓ i: The order of the primary unit (census sector) within the sub-strata
- ✓ ij: The order of the secondary unit (enumeration district), which means that the secondary unit has the order j within the primary unit of i order
- ✓ ijk: The tertiary unit (dwelling) which has three indices

Thus, the value of the given characteristic t (as for example the omissions of enumerating persons) will be expresses as follows:

- $t_{hi}$: The number of persons missed in the primary unit (census sector) in sub-strata h
- $t_{hij}$: The number of persons missed in the secondary unit (enumeration district) in sub-strata h
- $t_{hijk}$: The number of persons missed in the dwelling

The addition of the values of the sample units is symbolically expresses by Σ. One of the indices i, j, k is placed under the symbol of addition, shows at which stage and addition is meant. For example:

$\sum_k t_{hijk}$ : Total omissions of the dwellings selected in an enumeration district

$\sum_j t_{hijk}$ : Total omissions in enumeration districts selected in the census sector

$\sum_i t_{hijk}$ : Total omissions in census sector selected in the sample

The enumeration districts in the primary unit are represented by the symbol $M_{hi}$ and the ones selected in the sample are represented by the symbol $m_{hi}$

From the sample, we proceed to the population by stages, in the following manner:

### *3.6.1 List sample survey*

- Estimation of total enumeration district in sub-stratum h

$$\hat{t}_{hij} = \sum_k \frac{t_{hijk}}{p_{hijk}},$$

where $p_{hijk}$ is the dwelling selection probability in selection enumeration district

- Estimation of total in primary sampling unit

$$\hat{t}_{hi} = \frac{M_{hi}}{m_{hi}} \sum_j \hat{t}_{hij}$$

- Estimation of total sub-stratum from each one of the primary units in the sub-stratum separately:

$$\hat{T}_{hi} = \frac{1}{p_{hi}} \hat{t}_{hi}, \text{ (i=1,2)}$$

- Estimation of total sub-stratum from both primary units selected in the sub-stratum:

$$\hat{T}_h = \frac{1}{2} \sum_{i=1}^{2} \hat{T}_{hi}$$

The variation of this estimation of the sub-stratum (Krishmaih and Rao 1988). is:

$$V\left(\hat{T}_h\right) = \frac{1}{2}\left(\hat{T}_{h1} + \hat{T}_{h2}\right)^2$$

Now, the estimation of all sub-strata is expressed as follows:

$$\hat{T} = \sum_{h=1}^{4} \hat{T}_h ,$$

whereas the variation

$$V\left(\hat{T}\right) = \sum_{h=1}^{4} V\left(\hat{T}_h\right)$$

The coefficient of variation (%) is

$$CV\left(\hat{T}\right) = \frac{\sqrt{V\left(\hat{T}\right)}}{\hat{T}} \cdot 100$$

The results of the coverage survey are between 0.4% to 0.6% for the four censuses (1971, 1981, 1991, 2001).

### *3.6.2 Area sample survey*

The general formula for the estimation of T is expresses as follows:

$$\widehat{T} = \sum_{h} 1/2 \sum_{i} \frac{1}{p_{hi}} \frac{1}{m_{hi}} \sum_{j} \frac{t_{hij}}{p_{hij}}$$

The list and the area surveys supplement each other, as the first one refers to a sample of places of abode included in the Census whereas the second one refers to a sample of dwelling, which were missed. Consequently, both surveys give a sample, which is representative of total places of abode in country

## 4. PROCEDURE FOR SAMPLING MICRODATA

### 4.1 Sampling Fraction

For the purposes of IPUMS a random sample of 10% is selected from the private household that had been enumerated in the censuses of 1971, 1981, 1991, 2001. In detail:

i. The sampling fraction for the census of 1971 is $f = \frac{10}{25}$. This sampling fraction was chosen because the NSSG elaborated only the 25% of the private households for this census.
ii. The sampling fraction for the census of 1981 is $f = 1$. This sampling ratio was chosen because the NSSG elaborated only the 10% of the private households for this census and therefore all the available data are given to IPUMS.
iii. The sampling fraction for the censuses of 1991 and 2001 is $f = \frac{10}{100}$.

### 4.2 Geographical Stratification

For the selection of the sample from each census (excluding the census of 1981), the private households were stratified based on the NUTS III classification (Nomos or county).

In addition in each NUTSIII area the private households were also stratified based on the following criterion:

- Every municipality which has more than 20.000 inhabitants consists a separate stratum
- All the other municipalities in the specific NUTSIII area consist another stratum.

**4.3 Sampling procedure**

For each stratum *h,* $n_h$ private households are sampled from a total $N_h$ of private households in the specific stratum.

Systematic Random Sampling (Sharndal *et al* 1992) is applied for the selection of the $n_h$ households. In detail:

i. For each stratum, a sampling frame was created that includes all the $N_h$ households of the specific stratum. This frame (list) was sorted by the municipality of the households and the building blocks. In addition, a unique identification number is assigned in each household. This identification number is just the number of the line of the frame that this household is located, therefore the identification numbers take values from 1 up to $N_h$.

ii. Afterwards, the sampling interval was calculated as $\lambda = \frac{1}{f}$. Therefore, for the census of 1971 the sampling interval is $\lambda = 2{,}5$ and for the censuses of 1991, 2001 is $\lambda = 10$. Note that, the sampling interval $\lambda$ is not rounded to the nearest integer but is used as it is calculated**.**

iii. For each stratum we draw a random number $\rho$ in $(0, \rho]$.

iv. Based on the sampling interval λ and the random number ρ we produce the following series of number: $a_1 = \rho$, $a_2 = \rho + \lambda$, $a_3 = \rho + 2 \cdot \lambda, \ldots,$ $a_{n_h} = \rho + (n_h - 1) \cdot \lambda$.

In addition we transform the numbers $a_1, a_2, \ldots, a_{n_h}$ to a series of integer numbers by applying the following:

- $\beta_{hi} = a_{hi}$, when $a_{hi}$ is integer
- $\beta_{hi} = [a_{hi}] + 1$, when $a_{hi}$ is not an integer number, where [ ] denotes the integer part of the number.

v. The private households that have identification numbers one the number of the series should selected for data collection.

Note that, applying fractional interval method (Sharndal *et al* 1992) the sample size is controlled, because any element k will have chance $p_k = 1/\lambda = n/N$ of being chosen, and every possible sample will be exactly of size n.

Thus, it is clear that requiring λ to be positive integer will – in extreme situations – lead to sample sizes that may differ substantially from it is desired.

**4.3 Data provided to IPUMS**

Using the aforementioned sampling scheme, NSSG collect a sample from the Census data and this is provided to IPUMS project. In detail to number of private households and individuals that is given by NSSG to IPUMS projects are:

| Census Year | Sample of Households | Sample of Individuals |
|---|---|---|
| 1971 | 24,6153 | 845,473 |
| 1981 | 294,323 | 923,108 |
| 1991 | 320,391 | 969,407 |
| 2001 | 367,442 | 1,028,899 |

## 5. TESTING THE ACCURACY OF THE IPUMS SAMPLE

The NSSG in order to examine the accuracy of the sample that gives to IPUMS, uses the **Chi-Square test of goodness of fit** (Johnson and Bhattacharyya. 2001). This statistical test provides an estimate if a sample comes from a specific population.

In detail, using the Census of 2001, we determine the population by strata, 5-years age group and sex. These values are the population values (theoretical values) which are denoted by $E_i$.

In addition, using the 10% sample we estimate these population values by strata, 5-years age group and sex. These estimates are the observed values which are denoted by $O_i$.

Afterwards, we calculate the Chi-Square test for each 5-year age group by using the following formula

$$x^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

For the 5-years group, when $x^2 < x^2_{1-a/2}$ the sample of 10% provides accurate estimates for the population totals. In this analysis, confident level (a) is taken to be equal to 5%.

In our analysis, we concentrated on the age group 35-39. The Table below presents the results from the Census and the results from the sample by strata and sex

| Strata | Sex | Age Group | Census 2001 | | Sample | X^2 |
|---|---|---|---|---|---|---|
| | | | POPULATION | Ei=POP/10 | Oi | |
| 0103 | 1 | 35-39 | 1839 | 183.9 | 186 | 0.02398 |
| 0103 | 2 | 35-39 | 2042 | 204.2 | 224 | 1.919882 |
| 0199 | 1 | 35-39 | 5476 | 547.6 | 540 | 0.105478 |
| 0199 | 2 | 35-39 | 4585 | 458.5 | 470 | 0.288441 |
| 0301 | 1 | 35-39 | 749 | 74.9 | 74 | 0.010814 |
| 0301 | 2 | 35-39 | 787 | 78.7 | 94 | 2.97446 |
| : | : | : | : | : | : | : |
| : | : | : | : | : | : | : |
| : | : | : | : | : | : | : |
| : | : | : | : | : | : | : |
| A411 | 1 | 35-39 | 895 | 89.5 | 85 | 0.226257 |
| A411 | 2 | 35-39 | 999 | 99.9 | 107 | 0.504605 |
| A413 | 1 | 35-39 | 913 | 91.3 | 101 | 1.030559 |
| A413 | 2 | 35-39 | 923 | 92.3 | 81 | 1.383424 |
| A499 | 1 | 35-39 | 2588 | 258.8 | 265 | 0.148532 |
| A499 | 2 | 35-39 | 2360 | 236 | 243 | 0.207627 |
| | | | | | Total | 230.777 |

The result of Chi-Square test is 230.77 when the rejection area for this test is $x^2 > 272$. This means that the distribution of the sample is coming from the distribution of the population and therefore the sample could be used to produce accurate estimates for the population in the age group 35-39.

## 6. DISCUSSION

As the statistical information is obtained only from a sample of a units belonging to the population to be enumerated, sufficiently precise estimates will not be possible from small villages or communes. For this reason, the method of sample census data can be efficiently used only in those cases where information is needed in reference to relative large municipalities or groups of small and medium municipalities and communes.

## REFERENCES


Hogan H. and Wolter K. (1988). Measuring accurancy in a post-enumeration Survey. Survey Methodology, vol 14, pp 99-116.

Johnson R.A and Bhattacharyya G.K. (2001). Statistics: Principles and Methods. 4th Edition. John Wiley & Sons.

Marks E.S and Mauldin W.P. (1950) Response errors in census research. JASA, pp 424-438.

Kish L. (1995). Survey Sampling. Wiley Classics Library.
Kish L. (1979). Samples and Censuses. International Statistical Review, vol 147, pp 99-109.

Krishmaih R.P and Rao R.C. (1988). The techniques of Replicated or Interpenetrating samples. Handbook of Statistics, Vol 6. Elsevier Science Publishers P.V., 33-368.

Redfern P. (1987). A study of the Future of the Census of Population. Eurostat. ISBN 92-825-7429-6.

Sharndal C.E., Swensoon B. and Wretman J. (1992). Model Assisted Survey Sampling. Springer-Verlag, New York.