Contents

1 Project Description

1.1 The main objectives of the project

The sample described below has been used to collect data for the study "Social distinctions in modern Russia"(SDMR). The study was implemented in concert and with assistance from Finnish colleagues, representing the University of Tampere (Raimo Blom, Harri Melin and Jouko Nikula) and the Aleksanteri institute of Helsinki (Markku Kivinen). The project set out to span a wide range of various issues characterizing social differentiation in contemporary Russian society. In the context of the project the social difference was treated in a broad sense as a variety of human conditions (occupation, social mobility, property, income) as well as distinctions of attitudes towards civic life, politics, religion and other relevant issues. The logic of combining human condition variables and value questions aimed to explore the interaction between the two realms thereby testing numerous hypotheses related to their mutual influence. Another set of fairly innovative ideas covering various issues of social change might come from a possibility of a comparison between the results of contemporary data and the 1991 study "Social structure and class consciousness" (Coordinators E. Wright, M.Hout) of which Russia was a participant.

The set of issues accentuated in the project was the main rationale for the choice of sample design. The sample was to target the Russian population as a whole. It made little room for a detailed elaboration of territorial or ethnic issues that might require disproportionate or detailed representation

1.2 General principles of sample design

It is assumed that as any effective sample for any population the sample of the Russian Federation must be based on the following major principles:

1. The sample must allow for a reasonable equilibrium between the costs of the study and the precision of the data. It is obvious that in each case the equilibrium will be hitched to its own level of precision and costs. However, the possibility to control precision might make it possible to choose the optimal research strategy with concomitant conclusions about the exact sample size needed. The latter allows to limit the costs of the study and generate the highest possible precision of the data. In theory the bigger the sample, the more it precision it is capable to render. In practice each step towards more precision requires more investment and at some point drives the budget of the study in excess of the sum allocated for data collection. For example, while the sample of 10000 is certainly more preferable than the sample of 3000 thousand, the reduction of the sample error by 2.5% will in many cases not be contemplated as a reasonable balance for the tripling of field costs.
2. The sample design should be clear enough to allow for an easy replication in the future. While the design should be the same, the respondents should be different. In other words, the sample should not require consecutive visits to the same households unless it is elaborated for a panel study.
3. The sample must be well documented and open to the inspection of international experts. To ascertain the results of the study, the latter should be able to place reasonable requests for quality control procedures in the form of small-scale reliability studies based on a replica sample design for the main survey.

1.3 Sample constraints

The above-mentioned principles place a number of constraints on the actual sample design. Firstly, the sample must not invade any of remote areas or areas difficult to accede. In other words, it would be highly impractical and too costly to include into the sample the sparsely

populated areas in the Far North of Russia. Their inclusion into the sample would make little or no impact on the final study results (in the sample of 3000 they would be represented by no more than fifteen respondents), but in the same breath inflate the study's budget beyond reasonable costs.

The sample design should not require entry into dangerous or war-stricken areas. At present or in the near future it would hardly be feasible to incorporate areas in the Chechen, Ingush, Dagestan or Ossetian republics into any reasonable sample design. If a situation of civil unrest arises in another area, the sample design should be subject for quick change crossing it out from sample lists. In other words, given the cost, time and personnel safety requirements, the sample should take into account any dangerous developments in the subjects of the Russian Federation.

Table 1. Population in non-accessible areas excluded from the sample.[1]

Region	Population (in thousand)	Density (person per square meter)	Proportion of the population of the Russian Federation
Kaliningrad oblast	1072,0	428,0	0,72%
Chechen republic	813,0	66,8	0,55%
Ingush republic	309,0	59,1	0,21%
Adigei republic	450,0	49,6	0,30%
Dagestan	2074,0	59,1	1,39%
Northern Ossetia Republic (Alania)	665,0	83,0	0,45%
Kabarda-Balkar republic	790,0	32,3	0,53%
Sakhalin	713,1	8,1	0,48%
Taimir autonomous region	54,5	0,63	0,04%
Evenksky autonomous region	28,4	0,037	0,02%
The Kamchatka region	469,8	1,1	0,32%
Chukotsky autonomous region	155,7	0,21	0,10%
Yamalo-Nenetski autonomous region	495,0	0,66	0,33%
Khanti-Mansi autonomous region	358,7	1,24	0,24%
Tuva autonomous region	313,5	1,84	0,21%
Total	8761,7		5,88%

The sample must not require the selection of respondents in areas forbidden for visits of outsiders with no comfortable conditions for interviewing. Such are prisons, military units or hospitals. This rule poses problems because all of the named institutions comprise a significant part of the population in the working age, particularly males. The exclusion of this part of the population from the sample will inevitable be conducive to a reduction of the proportion of males in the sample.

Table 2. Population in institutions

[1] Demographicheski Ezhegodnik (The Demographic Yearbook of Russia), Moscow, Goskomstat of Russia, 1997, p.25-32

Institutions	Population (in thousand)	Proportion of the total population (%)
Russian armed forces	1300	0,8%
Forces of the Ministry of the Interior	250	0,2%
Inmates of prisons and correction camps	1000	0,7%
Inmates of hospitals and clinics	1788	1,21%
Total	4338	2,91%

According to the Red Star[2] newspaper men constitute close to 95% of the Russian army servicemen and Ministry of the Interior troops. Men make up 88% of all inmates of Russian prisons, correction camps and detention centers[3]. In hospitals the proportion is closer to that of the population. The above-mentioned facts will be taken into account in the process of comparing survey data and parameters of the population.

2 First Stage of Selection

2.1 Choice of PSUs

The design of the sample was based on the assumption that the overall sample size will be tantamount to about 1600 households or about 3000 respondents. The first stage of a multistage cluster sample develops the first level of classification of observation units in the population. The logic for the choice of a primary selection unit (PSU) is based on several vital conditions. Firstly, the PSUs must be as heterogeneous as possible. This condition implies that they must not be either too big or too small. Big units such as *oblasts* do not abide by the condition of homogeneity and therefore are unfit for this stage of selection. Electoral districts, on the other hand, are too small and while conforming to the first condition, they cannot be used as a fundamental unit of the selection process because any choice based on them would make the sample prohibitively expensive.

Secondly, the PSUs must be defined by clearcut geographical boundaries and described by primary statistical data. Access to the data must not be a problem as is usually the case with census units, well-described and relatively heterogeneous, but at the same time qualified as classified information by the State Committee for Statistics.

Thirdly, the number of PSUs should be over a thousand to contribute to the reduction of the sample error at the primary stage of the selection. It is obvious that the more PSUs are there for the first stage of selection, the more variance of the population parameters they will embrace.

Fourthly, the territory covered by each PSU should be reasonably limited to ensure access to any of its points by a team of interviewers. It is obvious that an *oblast* is too large to be covered by a small interviewer team. A proper selection of secondary sampling units in an *oblast* might

[2] Red Star, February 28, 1997

[3] Rossia v tsifrahk. Goskomstat Rossii. 1998. p. 120

thus be conducive to a heavy travel load for the usual team of two interviewers and therefore could complicate the process of data selection beyond reason.

The above arguments speak in favor of an administrative district (*rayon*) as the optimal choice for a PSU in the sample design of the Russian Federation. Firstly, *rayons* are comparatively small: their population might range from 50 to 300 thousand. However, it can be observed that smaller *rayon*s are usually found in rural areas and bigger ones – in urban centers. Therefore, a process of stratification into rural and urban *rayons* can eliminate a large part of the size-related variability of the PSUs. Secondly, *rayons* are fairly well described in statistical and other literature and the data on *rayons* are not classified. Thirdly, the number of *rayon*s in the Russian Federation is close to 2800 and that is a good basis for selection.

2.2 Source of primary data

The given sample design is based on the data of the 1989 All-Russia Census conducted by the State Committee of Statistics and the MicroCensus of 1994. According to the data the territory of Russia was divided into 2788 *rayon*s. The available statistical database has several limitations.

- It is evident that a lot of change has occurred in some parts of the Russian Federation since 1989. The change has made its most tangible impact upon the population of capital cities. To counter a possible bias, the latest update of the data (1996) by the State Committee of Statistics is used. The updated version of the data base is founded in the interim 1994 Microcensus conducted by the Committee.

- The data provide no evidence of the population of prisons, hospitals or army units situated on the territory of *rayons*.

- The data do not allow for a detailed examination of smaller settlements inside the *rayons*. In-depth information on territorial subdivisions, towns and villages had to be gotten by additional desktop research or through the offices of local research centers that sometimes had to undertake on-site inspection to generate up-to-date information on the local settlement structure.

Owing to the probability logic behind the design, the mentioned change would have to be quite dramatic to introduce any bias into the survey. It has to be borne in mind that whatever limitations of the data, they are the only reliable base for sample formation and will remain so beyond the year 2000 when the results of the next Census might be forthcoming.

The existing pattern of population distribution allows to divide all *rayons* into three types:

- Big city *rayons*, exemplified by Nizhni Novgorod

- *rayons* constituting territories with various settlement on them (towns and villages).

- urban *rayons* taken out of relevant territories subordinate to the larger administrative unit (oblast).

The last category poses a greater problem than the first two. In the given sample design the original statistical data were kept intact. The smaller urban settlements subordinate to the oblast were represented as pseudo-*rayon*s on a par with regular *rayon*s. The conversion was facilitated by the fact that the units in question are of the same average size as regular *rayon*s. The smaller urban type units situated in the vicinity of big cities (*PGT*) were clustered to produce a single pseudo-*rayon* comparable in size to smaller *rayon*s. This solution can be regarded as optimal from several viewpoints. First. The *rayon*s form fairly homogeneous strata favored by sampling theory. Second. The chosen way of grouping does not violate the original typology of data proposed by the Russian State Committee of Statistics. Thirdly, the solution makes it possible to drive the survey expenses down by collapsing the number of clusters. The overall number of *rayon*s thus manipulated is tantamount to 456.

2.3 Self-representing units

The sampling theory demands that self-representing units (SRU) should be chosen in line with two main principles:

1. The size of the unit. In almost every country there is one or several urban centers that are much greater than others. In Russia there are two such centers - Moscow and St. Petersburg. According to the State Committee of Statistics the population of Moscow is now close to 9 million, and the population of St.Petersburg -5 million.

2. The distinct social and cultural environment of the center offering more potential for research. Both Moscow and St. Petersburg are capital cities with an array of life styles distinct from the rest of the Russian Federation. Moscow is particularly specific and different from other Russian cities because it is traditionally a privileged city with a seat for many government offices and private companies. Currently over 90% of all Russian private enterprises are based in Moscow. It is a well-known fact that Moscow has gone a lot farther towards capitalism and private enterprise than any other Russian city.

When the number of strata is equal to 60-70, the sample provides for 5-6 of them as self-representing units. The SRUs cover about 20-25% of the population. When the number of strata is smaller, so is the number of SRUs. In the given sample design two SRUs are singled out.

Table 1. Characteristics of SRUs

SRU	Socio-economic zone	Population (in thousands)	Number of *rayon*s
Moscow	Central	8995	35
St.Petersburg	North-Western	5035	22
Total		14030	57

The SRUs are not regarded as separate strata and therefore are not divided into *rayon*s. Data collected in a SRU are representative. They can be taken out and analyzed separately. A survey in a SRU usually follows the pattern of any survey based on a representative sample. In Moscow, for instance, the sample was generated by a random choice of household telephone numbers. Since 95% of all Moscow households have access to telephone, it was possible to conduct a random selection using personal telephone number database. It may be said that the selection process in Moscow came as close to the pattern of simple random sampling as was possible.

Stratification

The theory of survey sampling assumes that a certain amount of information is known about primary sampling units. In the United States a sampler can make productive use of ample data on the geographical location of a PSU, its type (inner city, suburban or rural area), its population at the previous census, its rate of population change, the proportion of the population employed in manufacturing, the proportion of non-white population resident in the area. As was said before, the Russian statistical bureaus do not to provide as much information. The data on the PSUs are limited to the description of its geographical boundaries, its type and its population size.

Nevertheless, even the available data suffice to launch the process of stratification. The latter implies the division of the population into subpopulations, or strata, based on the supplementary statistical data and creating separate samples from each of the chosen areas. Stratification is essential in Russia that abounds in sparsely populated areas. Under the existing circumstances stratification allows to control the size of the sample and at the same time ensure the *epsem* (equal probability of selection) principle for each individual in the population. In the present design we propose to make the strata sample size proportional to the population sample size. In other words, we apply the *uniform sampling fraction* principle

As is proven by the sampling theory stratification makes a sample no less precise than a simple random sample of the same population.[4] The number of the strata is usually determined on the basis of several major criteria. First. The division of the sample into strata should end up with approximately 20 interviews per interviewer and in the conditions when two interviewers are employed - with 40 interviews per strata. The engagement of two interviewers is paramount since a good deal of travel is involved. Besides the conduct of interviews by two employees will allow to utilize sophisticated quality control measuring when the measure of interviewer variance might serve as an indication of possible error generated by the process of data collection. A larger than 20 number of interviews may cause an overload for each interviewer and consequently be detrimental to the quality of the data thus obtained. On the other hand, a

[4] Kalton Graham. Introduction to survey sampling. A Sage University paper.1983. p.20.

smaller number of interviews per interviewer is likely to dwarf his or her income and leave the most qualified field staff dissatisfied with the conditions of their involvement in the project[5]. Fortunately, there was a chance to bypass this problem in the present study by way of forming not one but several locally-based teams of interviewers in every region.

Second, each stratum should form a relatively homogenous unit with little variance over demographic or cultural characteristics.[6] Since we do not have any detailed information on the units, we follow up the following principles of aggregating PSUs into a stratum:

- The PSUs should be approximately similar in size. The size of PSUs is provided by the available statistics from the data of the last Census.

- The PSUs must be of the same type. We divide PSUs into urban (more than 66% of the population is urban), mixed (from 33% to 66% is urban) and rural (less than 33% of the population is urban).

- The PSUs must be situated in geographical proximity to each other. Grouping rayons from the same or adjacent socio-economic zone ensures this criterion.[7] In the majority cases the *rayon*s are drawn from the neighboring regions (*oblasts*). The available data indicate that this degree of proximity is highly likely to ensure similar cultural characteristics of the PSUs. The characteristics in question comprise such important indicators as family size, life style and consumption patterns.

Third. The number of strata should not exceed 50 units. Any other option might require too large a workforce and boost expenses. While there is always a possibility to employ additional staff, the number of qualified interviews cannot be quickly increased: "The interviewer skills constitute an expensive and important asset. This places a premium on the continued

[5] The number of interviews conducted by one interviewer within one survey should be limited. In a number of methodological experiments it has been revealed that an interviewer if faced with more than fourty interviews develops undergoes a process of "education", He often feels that he is increasingly qualified to interpret the arguments he gets from the respondents. He or she develops an attitude described in the methodology of social research as "selective listening". (Elizabeth Noelle. Umfragen in der massengesellshaft. Munchen. 1971.) In other words, as a result of an overload the interviewer is less and less inclined to reduce himself to recording the answers from the respondent and and is more and more likely to provide his own interpretations to the responses he gets. The problem can be avoided if the interviewer's load is limited, ranging from 20 to 30 interviews in one survey.

[6] In Survey Sampling by L.Kish the problem is formulated in the following fashion: "The aim is to from strata within which the sampling units are relatively homogeneous in survey variables. Their variances are reduced to the extent that the variation among sampling units within the strata is less than their variation in the entire population. Hence, we strive to increase and maximize the homogeneity of the sampling units within strata. For a given population of sampling units, this is equivalent to increasing the differences, or heterogeneity among the means of the strata." (L.Kish. Survey sampling. John Wiley and Sons, New York, 1965, p. 100).

[7] According to the Russian Statistical Agency, the territory of the Russian Federation is divided into 11 socio-economic zones: Northern, North-Western, Central, Central Black-Earth, Volgo-Vyatsky (Northern Volga basin), Povolzhski (Volga river basin), Northern Caucasian, Ural, West Siberian, East Siberian, Far Eastern

employment of the interviewers for several years at least. After the initial training, most of the further training and instructions for specific surveys are handled with mailed materials."[8]

We assumed that the present sample design would require no more than 24 strata - an arrangement fitting in with all above conditions. The average size of a stratum might range from 3 to 6 million inhabitants. The rural strata may be smaller as a result of a response to large-scale depopulation of many rural areas.

2.4 Standard error

One important aspect of our study is to provide information on the precision of the estimates made from the survey data. Since the design does not require the questioning of all possible respondents in Russia, but only a selection of them, we can expect to obtain somewhat different results of randomizing procedures produced a different sample each time. What should estimated , therefore, is the degree to which the results that are obtained are subject to this variability, called sampling error. The sample design proposed is a complex one involving the selection of clusters of households by the three-stage sampling procedure which will be described lower. This factor has the effect of increasing sampling error beyond that expected from a simple random sample of the same number of elements. For samples of this kind special procedures are needed in order to estimate sampling error. This procedures differ from the application of simple formulas such as pq/n (p and q are the distribution in the sample, n is the number of respondents) that assume simple random sample. The method adopted here is based on the Taylor series approximation to the variance of a ratio mean. It uses the differences in sample results found between pairs of primary sampling units selected within each of the many strata.[9] The computations is usually carried out using the WestVar software created by SPSS corporation.

$$(1) \qquad Dh = (Yh1 - Yh2) - r\,(xh1 - xh2)$$

where r is the ratio y/x being estimated based on all respondents in the sample. For the means and proportion estimated for this study, x is simply the number of cases in the sample, and y is the value of the variable whose mean is being estimated (for proportions, the value of y is either 1 or 0). Subscripted values of x and y refer to the values within a specific primary sampling unit (1 or 2) within stratum h. Note from equation 1 that we are interested in the differences between the values of the x's and respectively y's within stratum h. Once the value of Dh has been computed for each stratum, it is squared, summed across strata, and the sum is divided by the square of the sample size, yielding the estimated variance of the ratio r.

$$(2) \qquad \text{var}(r) = 1/x^2 \; E \; D^2_h$$

The standard error is the square root of the variance computed from equation 2.

[8] Kish L. Survey sampling. John Wiley and Sons. Inc. New York. 1965. p. 366.

[9] Computation of variances of this paired selection method is described in Kish, Survey Sampling, pp. 190-195; see equation 6.4.8. This method has been widely used, discussed, and recommended.

Using these procedures and the computer program we can generate in a mechanical way, an estimated standard error for each of the percentages reported in the tables of study results.

The above variance is multiplied by standard deviations of the sampling distributions known as the standard error. Let us denote the standard deviation of a simple random sample as y_0 (with the subscript 0 to indicate simple random sampling), its standard error by SE (y_0) and the square of the standard error, the variance of y_0 by V(y_0) Most standard sampling error can be presented in terms of variances rather than standard errors. The variance of a sample mean in a simple random sample of size n is embodied in

$$(3)\ V(y_0) = N\text{-}n/n\text{-}1 * \delta^2/n$$

or converted from the previous formula

$$(4)\ V(y_0) = (N\text{-}n/N)\ S^2/n = (1\text{-}f)\ S^2/n$$

where n/N is the sampling fraction (f)

According to the formulae above the V(y_0) is dependent on three major factors:

1. *The finite population correction* (fpc) shown in the formula (N-n)/(N-1). Obviously it is negligible in the case when the population is large. For the sample of the Russian population it will be tantamount to (148000000-3200)* (148000000-1) = .999. In other words fpc testifies to the paradoxical fact that the larger the population the less important is the size of the fraction that the sample is equal to.

2. *The sample size*. As can be seen from the formula the larger the sample, the smaller is V(y_0). In other words, for larger populations, such as the population of Russia or, for that matter, the population of a big city, the size is much more important in determining the precision of survey results. The sample of 3200 drawn from the population of a country renders the same results as the sample drawn from the population of a big city.

3. *The variance (S^2)*. If all elements of the population choose to respond in the same fashion, the standard error will be negligible. If, on the other hand, they differ, there is a probability that the population mean will greatly differ from the sample mean. Hence to determine the standard error the mean of the sample should be compared with the mean of the population. The advantage of equation 4 is that S^2 has an unbiased estimator in the form of $s^2 = \Sigma (y_i - y_0)/(n-1)$ Therefrom comes

$$(5)\ \text{se}(y_0) = (1\text{-}f) \sqrt{s^2/n}$$

with lower case letters indicating sample esimators.

When the standard errors is estimated, a confidence interval should be established. If the sample is large, the 95% interval for Y is y_0 +|- 1.96 se (y_0). The 1.96 is taken from a table of the normal distribution, where 95% of the normal distribution falls within 1.96 standard deviations around the distributions means. So the final formula for estimating the sample error would consist of the estimation of the effect of clustering and the evaluation of standard error as such.

$$(6)\ se_s = 1.96\sqrt{\text{var}(r)/n \quad (1-f)s^2/n}$$

It is evident that we have no means to estimate var (r) prior to the survey itself. We can, however, estimate the maximum standard error (50/50 distribution) for the three proposed samples (2600 without an account for the design effect. In this case the sample error might be equal to 1,92.

As our experience shows the design effect raises the sample error 2.1-2.2 times in the samples comprising from 2000 to 3000 respondents. In other words, the sample size of 2600 will yield the sample error of approximately 5%.

2.5 *Rayon* selection[10]

Once the list of primary areas in each stratum is complete, one is selected from each stratum. The selection was carried out using the procedure described L.Kish.[11] Prior to selection, the *rayon*s were grouped according to their size and for mixed stratum – also according to the proportion of the urban population. Each rayon received a number and after that a random selection of one single PSU was taken.

3 Second Stage of Selection

3.1 Choosing the secondary sampling units

The second stage consists in the selection of secondary selection units (SSUs). The sample frame provides several options for a SSU in various areas. As the PSUs the SSUs must be of an approximately equal size. Census districts can be used as the SSUs, however as has been mentioned above, the information about them is scarce. The second strategy of selection can be based on electoral districts. They are small enough and can be served by one or two interviewers. In addition there are variegated data about them including a list of housing blocs and lists of residents. The latter are, however, quite frequently out-dated or unavailable. We assume that an original strategy of selection should be employed in each PSUs depending on access to primary data about SSUs. In some census districts might be used, in others - electoral districts or postal areas.

Let us analyze the procedure of selecting a SSU and within a PSU. In an imagined PSU there 3 towns, 4 urban type settlements and 12 villages. We choose two SSUs in the given PSU.

[10] With SRUs left out the base for strata formation incorporates **2492** *rayon*s and pseudo-*rayon*s.

[11] L.Kish. Survey Sampling. New York.John Wiley and Sons. p. 230-231

Table 5. Process of SSU selection

Selection of SSUs. (ED - electoral district)Settlement	SSU	Population (thousands)	Stratum population (thousands)	Cumulative population
Town 1	ED-1	2,2	9,0	2,2
Town 1	ED-2	2,2		4,4
Town 1	ED-3	2,1		6,5
Town 1	ED-4	1,3		7,8
Town 1	ED-5	1,2		9,0
Town 2	ED-1	2,0	6,7	11,0
Town 2	ED-2	1,8		12,8
Town 2	ED-3	1,5		14,3
Town 2	ED-4	1,4		15,7
Town 3	ED-1	1,8	3,5	17,5
Town 3	ED-2	1,7		19,2
Village 1		1,7	4,2	20,9
Village 2		0,8		21,7
Village 3		0,8		22,5
Village 4		0,5		23,0
Village 5		0,4		23,4
Village 6		0,1		23,5

In line with the PPS principle electoral districts are grouped into strata according to the type of settlement and size. These are not the only possible grouping criteria. In *rayon*s with mixed ethnic composition ethnicity may also be brought in and, correspondingly, each ED is categorized as having either ethnically homogeneous or heterogeneous population.

To systematically select the SSUs we need to know the interval to abide in the process of selection. The interval is equal to n/2 where n is the total population of the PSU. In our case the interval is equal to 11,75

The process of selection consists in adding 11,75 to the number between 0 and 10 generated by the random number generator. Let the number be equal to 5. So the first SSU selected is closer to five on a cumulative scale in the right-hand column. This is Town 1-ED2. The second SSU chosen is the one pinpointed through the following calculation.

(7) Second SSU = 5 + 11,75=16,75

The closest SSU to this point on a cumulative population scale is Town 3-ED1. It is then selected as the second SSU to be visited by interviewers.

3.2 Number of interviewers in a SSU

Our next step will consist in determining the number of interviews for each ED selected. It is paramount that each households chosen should be characterized by the same probability of being selected as all other households of the area. The probability results from the probability of selection determined for the *rayon* (p1), the probability of selection for the ED (p2) and the probability of selection of a given ED resident (p3).

P1 is a ratio of the given *rayon* population and the population of a given stratum. P2 is a ratio of the N population of a given ED to the population of the *rayon* multiplied by 2. P3 is equal to the number of interviews held in a given ED divided by its population. At the same time it is the ratio of the sample size (number of households) to the Population of Russia (total number of households in Russia) divided by the P1*P2. In other words, the number of households in a *rayon* will be computed on the basis of the following formula:

$$(8)\ N = Ped * p/p_1 * p_2$$

where N is the number of interviews, P_{ed} is the population of a given ED, P is the probability of selection for each household of the Russian Federation, P_1 is probability of selection of a given stratum and P_2 is the probability of selection of a given *rayon*. Let us imagine that the population of the stratum is equal to 4 mln or 1,1 million households.

Table 6. The number of households in each ED.

Settlement	SSU	Population	p1	p2	N
Town 1	ED-1	2200	0,042	0,00588	18
Town 3	ED-3	1800	0,042	0,00588	18

On the average the number of households per electoral district is equal to 18. It implies holding about 100 interviewers with individual respondents.

4 Third Stage of Selection: Choosing Respondents

There are several ways for choosing respondents. For many years Russian pollsters relied on electoral lists as a source of names and addresses. They are conveniently organized and in case of need make it possible to effect systematic selection of a preset number of respondents. However recent changes in the electoral system allowing citizens to abstain from voting decreased the quality of the lists. It is no longer obligatory for electoral committees to doggedly pursue every respondent if he or she chooses to skip their civic duties. In addition in many cases a more mobile population, especially in big cities, has attenuated a link between the formally registered and real place of residence. Quite frequently squeezed for more income urban dwellers lease their apartments to strangers who are not to be found in the relevant electoral lists. One more fault of the lists is that they contain only citizens of over eighteen.

The present design provides a more reliable alternative to the electoral lists. While data contained in the electoral lists is increasingly unavailable (frequently classified) data on the housing blocs situated in the district is no secret. As a rule, in urban areas an electoral district contains from 5 to 10 housing blocs. Prior to the survey an interviewer can easily list bloc apartments. Actually each apartment can be regarded and is a household unit. When they are listed down, a selection of households can be done.

In rural areas the interviewers have relied and will continue to rely on household books. The book lists all households and their members residing in a given village. It opens the possibility of a systematic selection that would start with the random figure. In practice it is often required

that the data of the household books should be checked and rechecked against someone's inside knowledge. It is not too infrequent that members of the family listed as living in the village are if fact absent. It is often the case with old women who, shunning the fatigue of village life in winter, spend winter months with their relatives in a city.

5 Analysis and Results

5.1 Comparison with Census data

Below is the comparison between some of the demographic characteristics of the achieved sample and those of the operationalized population. The operationalized population is treated as the data coming from the Microcensus of 1994 with relevant subtractions of the groups excluded from the study. It will be assumed that the Census provided what can be termed as "objective data". However, while intended to be a good sample of the population, the one million Census came under harsh criticism from international experts. The samplers paid by the World bank claimed that the MicroCensus can hardly be qualified as a representative study since its sample followed in the footsteps of Soviet tradition and chose respondents in line with industry division rather than through household selection. The only alternative to the Microcensus is the 1989 Census data, but, as has been stated above, they are largely out of date. To counter the above mentioned problems, we shall test the study results not only against the data, but also against another dataset generated by a large-scale RLMS project based on a random probability sample of about 10000 respondents.[12]

The tests of sampling error carried out here compared the proportion of the SMDR and RLMS with a specific demographic proportion of the census data. The proportions are used to calculate a test statistic Z on the basis of the formula:

$$(9) Z = (Ps - P) / \sqrt{P(1-P)/n}$$

Where
Ps=the proportion from the sample
P=the proportion form the population
N=sample size

The null hypothesis is that the differences in the sample can be attributed to sample error. The final columns, then, are the p-values: the probability of obtaining a value of Z at least as extreme under the null hypothesis using a Gaussian distribuiton (normal).

[12] RLMS (Russian Longitudinal Monitoring Study) is a study gauging income and health characteristics of the Russian population. According to the sample design of RLMS the population is divided into thirty strata and in each stratum numerous PSUs three PSUs are selected. The RLMS targets households rather than individual respondents. Inside a household all respondents are questioned.

Table 7. National comparisons

Parameter	Category	Statistic (%)	SDMR(%)	RLMS (%)	P-value SDMR	P-value RLMS
Gender	Male	46,0	42,7	43,0	0,02	0,00
	Female	54,0	57,3	57,0	0,02	0,00
Age	20-29 years	18,9	19,0	18,4	0,90	0,13
	30-39 years	22,3	20,0	20,6	0,06	0,00
	40-49 years	21,2	23,9	20,3	0,02	0,02
	50-59 years	14,0	13,7	14,4	0,66	0,10
	60 years and older	23,7	23,4	26,2	0,72	0,00
Education	Higher	16,1	18,6	17,4	0,02	0,00
	Unfinished higher	2,1	3,1	0,8	0,02	0,00
	Secondary	54,4	58,8	52,0	0,00	0,00
	Lower	27,2	19,2	29,8	0,00	0,00

The P-values indicate that in most categories the precision of the SDMR was on a par with the RLMS – a much bigger and more expensive survey. The SDMR did better in representing the younger category (20-29 years) and the oldest one (50 years and older). In most basic categories the proportions stay within the 5% margin of error with a 95% confidence interval. There is only one category dramatically undersampled – the category with a low level of education. A closer inspection of the data shows that the undersampling of the category might be linked with the process of operationalization of the population e.g. the exclusion of certain areas where residents with an unfinished secondary education constitute a large part of the population. These are the Caucasian republics, qualified as either inaccessible or hazardous. Indirectly the hypothesis is substantiated by available statistical data: while in Russia as a whole 73,1% of the population living in urban areas, in the Chechen republic it is 34,7%, in the Ingush republic – 41,8%, in Dagestan – 41,8%. A larger rural population tends to be correlated with a larger category an unfinished secondary education.

5.2 Non-response

Non-response is another possible source of bias in any survey. Let us estimate the overall influence of the non-response upon the quality of the data. First, let us suppose that the overall mean for the population would be as follows:

(10) $Y=PrYr+PnYn$

where:

Y=mean for the population

Pr=proportion of responses

Pn=proportion of non-response

Yr=mean for the responses

Yn=mean for the non-response.

The proportion of responses and non-responses add up to 1.

The estimate for the mean of the non-response stratum of the population is as follows:

(11) $Yr-Y= Yr-(PrYr+PnYn) =Yr(1-Pr)-PrYn =Pn(Yr-Yn)$.

In our case, the total non-response rate amounted to 0,21, so the formula will look as:

Non-response bias=0,21(Yr-Yn).

In other words, if, for instance, the average difference of income between responding part of the population and non-response part is equal to 100 rubles, the total impact upon the mean

income would be equal to 21 rubles. In our case the average income is about 400 rubles. The overall impact of non-response would not be conducive to any serious bias.

The problem of non-response is traditionally more acute in Moscow. The losses caused by non-responding residents in Moscow are equal to 28% - a significantly bigger proportion than in the Russian provinces.

6 Conclusion

When a survey is finished, the obvious question of how a better representation could have been achieved is always at the forefront of a researcher's mind. It is obvious that the sample design could have been greatly enhanced by an increase of the sample as well as by more PSUs chosen within every stratum. A better result could have been attained if respondents were paid: such is the policy of the RLMS organizers. However, in all of these instances the budget of the survey would have been bound to dramatically rise. In our view, the design of the sample strikes an equilibrium between the required precision and the chances of collecting a set of representative data characterizing the Russian society at a crucial period of its history.

7 APPENDIX 1. PSUs in the sample

	PSU	Socio-economic zone	Number of households selected	Number of respondents
1	Moscow	Central	182	300
2	S.Petersburg	Central Western	130	301
3	Kurkinsky rayon	Central	51	108
4	Torzhokski rayon	Central	50	101
5	Nelidovo	Central	71	120
6	Eletz	Central Black Earth	69	114
7	Khasanski rayon	East Siberian	52	77
8	Bogotol	East Siberian	56	118
9	Uyarski rayon	East Siberian	36	67
10	Novopokrovski rayon	North Caucasian	47	93
11	Bataisk	North Caucasian	63	132
12	Podporozhie	North Western	48	100
13	Kirillov rayon	Northern	28	57
14	Tunkansi rayon	Northern	33	64
15	Tukaevski rayon	Povolzhski (Volga)	54	81
16	Krasnopartizanski rayon	Povolzhski (Volga)	50	76
17	Balashov	Povolzhski (Volga)	57	79
18	Rtischevo	Povolzhski (Volga)	55	73
19	Novosergeyevski rayon	Ural	43	83
20	Buzuluk rayon	Ural	41	69
21	Blagovarski rayon	Ural	40	90
22	Chelyabinsk	Ural	48	80
23	Emanzhelinsk	Ural	57	104
24	Nizhni Novgorod	Volgo-Vyatski	56	111
25	Kuybishev	West Siberian	50	96
26	Tumen	West Siberian	49	110
	Total		1516	2804

8 APPENDIX 1. Nonresponse

	PSU	Number of respondents	Number of refusals	Number of not-at-homes	Other non-response
1	Moscow	300	68	39	11
2	S.Petersburg	301	3	12	6
3	Kurkinsky rayon	108	0	8	1
4	Torzhokski rayon	101	5	16	2
5	Nelidovo	120	12	15	7
6	Eletz	114	7	4	4
7	Khasanski rayon[13]	77	4	21	1
8	Bogotol	118	2	0	0
9	Uyarski rayon	67	0	2	0
10	Novopokrovski rayon	93	11	13	5
11	Bataisk	132	1	4	2
12	Podporozhie	100	0	9	0
13	Kirillov rayon	57	1	3	0
14	Tunkansi rayon	64	4	3	3
15	Tukaevski rayon	81	9	26	8
16	Krasnopartizanski rayon	76	18	12	4
17	Balashov	79	11	23	4
18	Rtischevo	73	19	24	3
19	Novosergeyevski rayon	83	7	3	1
20	Buzuluk rayon	69	3	13	2
21	Blagovarski rayon	90	0	7	1
22	Chelyabinsk	80	7	15	2
23	Emanzhelinsk	104	8	7	0
24	Nizhni Novgorod	111	0	11	1
25	Kuybishev	96	0	6	2
26	Tumen	110	2	6	0
	Total	2804	202	302	70

Available data show that the rate of non-response for the SDMR survey is tantamount to 20,5%. The non-response is made of refusals (7,2%), not-at-homes (10,8%) and other losses (sickness, etc.=2,5%). In terms of refusals the survey came out with a good hit rate. The not-at-homes tended to constitute a large part of the population that in itself is a basic characteristic of modern Russian society: massive unemployment forces many Russians to become temporary dwellers of regions other than their own.

[13] In some PSUs interviewers faced nonresponse caused by a large absentee rate. For instance, in the Khasan rayon a large number of men were sacked from local enterprise and to earn a living had to get jobs as fishermen on the ships of the Far Eastern fleet.