Developments in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 19, Ljubljana: FDV, 2003

New Sampling Designs and the Quality of Data

Juergen H.P. Hoffmeyer-Zlotnik[1]

Abstract

Classical random-route sampling designs were continuously modified in order to reduce costs on the one side and as a reaction to a new type of full time interviewer on other side. The modifications have led to a situation where a random route sample is no longer a probability sample. Big social- and market-research institutes combine the random walk with a quota approach.

In this paper three different types of random-route sampling designs are compared: a classical approach with a controlled random walk and in a different step the interviewing of a fixed gross N of interviewees, a random walk defined by the number of net N completed interviews, and a random walk combined with a quota approach. The different types of sampling are discussed and the results of fieldwork demonstrated.

1 Introduction

A sample is a selection of units from a defined population. In survey research a sample is a selection of persons or households from the resident population in private households. Survey researchers expect that the selected persons or households are representative for the population of interest. The goal is to get an unbiased, representative sample of the population of interest which can be used to infer the behavior/attitudes of the target population.

Sampling methods in survey research are classified as either probability or non-probability. In probability samples, each member of the population has a known non-zero probability of being selected. Probability methods include random sampling, systematic sampling, and stratified sampling, or a mix of all these. In non-probability sampling, members are selected from the population in some non-random manner. In survey research this is quota sampling. The advantage of probability sampling is that sampling error can be calculated. Sampling error is the degree to which a sample might differ from the population. When generalizing to the population, results are reported plus or minus the sampling error. In non-

[1] ZUMA, Mannheim, Germany; Hoffmeyer-Zlotnik@zuma-mannheim.de

probability sampling, the degree to which the sample differs from the population remains unknown.

This article concentrates on random route sampling techniques. First, different methods or techniques of random sampling are discussed. Special attention is given to a new variation called “random route plus”. Second, data collected by these different methods of random sampling are compared to each other.

2 Random Route Sampling Methods

The purest form of probability sampling is a random sample. Each member of the population has a known chance of being selected. Random sampling in large populations is only possible if every member of the population can be identified.

Therefore, national samples are based on multi level selection processes. Multi level selection processes are a combination of random or/and systematic and/or stratified probability samples at different levels: stratified selection of sampling units, systematic selection of households by random walk, random selection of one person per household by Kish-table.

The sampling design that is mostly used for random route sampling in Germany is developed by ADM (Association of German Market and Social Researchers).

In a typical sampling process the first step is the selection of the spatial level: the sampling units. A sampling unit is a delimited small area for random walk in which as a minimum of 200 persons of the target population are living. In German ADM-design these sampling units are the "voting districts" for national elections (administrative subunits of the constituancies) in which as a maximum about 2.000 eligible voters are living. The sampling units are organized in national networks about 258 "voting districts".

The second step is a random walk per "voting district". Starting from an address which was selected ad random interviewers walk from house to house on a prescribed route. During this walk dwelling units with households are listed in prescribed steps: for example every 10^{th} unit (here: household) is listed. On this level many variations of walking and listing addresses are possible. You can separate or combine listing of addresses and interviewing of target persons, you can list a limited or unlimited number of addresses.

The third step of sampling, executed on the level of household, is the selection of the target person by Kish-table. This is a simple random sample: Each person of the household if he or she is a member of the population of interest has an equal and known chance of being selected.

3 Different models of Random Route

Especially in steps two and three the sampling process can be simplified and costs of data collection can be reduced.

Table 1: Different Models of Random Route Sampling.

	Controlled sampling	Uncontrolled sampling	RR+Quota sampling
1st step:	voting district	voting district	voting d.
2nd. step:			
• selection of addresses for interviewing	address listing separated from interviewing	address listing integrated with interviewing	
• requirement	gross number of addresses defined, net number of interviews undefined	net number of interviews defined, gross number of addresses undefined	
• substitution of dropouts	no	yes	yes
• protocol:	yes	no	no
3rd. Step:	Kish-table	Kish-table	Quota

Table 1 gives three forms of random route designs where the first level of sampling, the selection of the sampling points, is the same for all three methods.

By the controlled sampling households are listed by interviewer "A" walking the route according to the prescription. The listed households are the frame for the local sample of households. Then interviewer B gets N addresses of the sampled households and is asked to contact them selecting the target person by Kish-table. After the target person of a household is defined the interviewer can contact him or her for interviewing. In this version of sampling design the interviewer has the order to complete a maximum of interviews from the given N persons (one per household) belonging to the target sample. If interviewer B realizes only a small number of interviews, then interviewer C is sent to the sampling point to try again to complete a maximum of missing interviews.

A first modification of this design combines the listing of households with interviewing the target person. Sampling on the second level is no longer done in

two different steps. Now interviewer A accomplishes the listing of the households, selects the target persons in every N^{th} household according to the Kish-table and completes the interview. Drop a separate walk for listing households will save 20% of expenses. But the control of the interviewer's walk and listing of households is hardly possible any longer.

A second modification of the random walk is described as "uncontrolled sampling" in Table 1. Here the listing of exact addresses as well as the definition of the sample size is dropped. The interviewer can contact, e.g., each third address as long as he or she needs to complete the given number of interviews. The walk is as long as an interviewer needs to realize the given number of (net) N interviews. In this modification of random route the gross of the sample is no longer defined. Instead of the gross the net of needed interviews is given. In comparison to the first modification this second modification saves 30% of expenses. But absolutely no control of the interviewers is possible. And the interviewers contact persons who are easy to reach in the mid of the day like housewives and pensioners. So we get an over-sampling of young women and old men, of young households with children and of unemployed persons. By order the interviewers should use the Kish-table for selection of the target persons. But they substitute the target person by a reachable person who is willing to be interviewed. The result is a large bias in the distribution of contacted persons. With this modification of the sampling process, random route sampling turns from probability to non-probability sampling. The sampling error can not be calculated.

The third modification is the newest creation of German market research institutes: the so called "random route plus quota sample". This modification is described in the last column of Table 1. The selection of the target person within the household by Kish-table is replaced by quota. Here a random walk with a given number of net N interviews is combined with quota sampling.

In quota sampling the selection of respondents is non-random. In quota sampling, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then it is left to the interviewer to select the persons from each segment based on a specified proportion (often not in proportion to the local population): The interviewer is told that the completed quota of questionnaires must include a certain number of males and of females and a certain number from specified age groups. So, the sample is chosen by the interviewer from the resident population living in the streets he/she passes by on the "random" walk. It is this step which makes the technique one of non-probability sampling. Such a technique of sampling can be open to a great deal of abuse resulting in bias. It is the interviewer who defines selection criteria for the quota during walk, by the subjective form of acquisition, and by a specific wise of access.

A quota sample is cheap because it eliminates repeated calls to interview a person who may not be at home. In practice this modification does not save any more expenses because the interviewer does not use the Kish-table even in modification two (in Table 1). Differing from the second modification this new

modification pretends a better distribution of contacted persons because the data are self-weighted by those variables one can control by Micro-Census information (gender, age, and education). On his/her prescribed walk the interviewer can contact each person he/she will meet if this person is within the quota. So the interviewer is allowed to do interviews during the whole day. In combination with a quota technique at the last step this sampling procedure is a real non-probability sample. The sampling error can not be calculated.

4 Random Route Plus (Quota)

The first sample drawn by the so called "Random Route Plus" method in Germany is well documented. Thus, the field process can be reconstructed (see Table 2).

Table 2: Random Route Plus: Coverage of a sample.

	∅/point	%	range
n of listed households	45	100	26 - 70
non-contact	8	19	1 - 14
out of sample	15	35	
• no person of sampled population	10	22	3 - 22
• no German language	1	2	0 - 4
• Quota just filled	5	11	0 - 12
contacted households in sample	22	100	
eligible, non interview	14	58	
• household level refusal	5	22	1 - 12
• respondent not at home	1	2	
• respondent refusal	8	34	1 - 19
realized interviews/point	8	37	3 - 10

On average, 45 households were listed in each sampling point. In 8 of these households nobody was at home. These “no contacts” were excluded from the sample before a second contact could start. Of the remaining 37 households 15 households were defined as neutral losses. Furthermore, losses resulting from the required quota characteristics also are added to the neutral losses. These losses resulted from the fact that, independent of the willingness to participate, a

contacted person had to be defined as a loss if a field in the matrix given by the quota already was filled with interviews.

Another loss of 13 addresses per sample point result from the so called systematic loss produced by refusals at the level of the households as well as of the individuals. Thus, in an average sampling point 8 interviews are completed out of 45 households which are listed altogether. The interviewers were asked to complete 10 interviews per sampling point. The quota characteristics stratified the sample into six groups given by 3 age groups (18-26, 27-39, 40-55) for males and females. In the two groups of younger people only one interview per group was allowed, in each of the other four groups two interviews were required.

5 Selection of target persons in a sampling point

Selection of the target persons in the sampling points was done by the interviewers. Therefore, they got the table of the quota characteristics (ten interviews allocated in six groups defined by age and gender), and they got a starting address and the prescription of the walk (direction of walk, interval of addresses, who should be contacted, and a decisions guide for the walk if another street is crossing).

In theory the quota characteristics should be given individually for each sampling point. But in practice one quota characteristic is given for all sampling points. That is the first mistake. The second one is that the interviewers will not list all persons per household selecting those persons who are marked by quota characteristic at random. The interviewers like to contact one household only once. And they are interviewing those people who are available and are willing to give an interview.

In my view this sample can not be better or even worse than the walk from house to house interviewing all those persons who are available and are willing but not defined by quota. By both strategies the interviewer is interviewing easy to reach people. The quota here appears as a strategy of weighting. During the last few months of 2001 a national survey in Germany used this sample design mixing up random route and quota - the so called "Random Route Plus" (quota). Now we can answer the question about what happened in the fieldwork. For demonstration in Table 3 two sampling points were picked up at random.

Table 3: Interviews proceed per sampling point. Example of two sampling points.

person	1	2	3	4	5	6	interview
quota	1	1	2	2	2	2	no
point A:	0	1	1	0	0	N	
	0	0	0	0	1	1	N
	0	1	0	0	0	0	1
	1	0	0	0	0	0	N
	0	0	1	0	0	0	N
	1	0	0	0	0	0	2
	0	0	1	0	0	0	N
	0	0	1	0	0	0	3
	0	0	0	0	1	0	N
	0	1	0	0	0	0	4
	0	0	0	0	0	1	5
	0	0	0	1	0	0	6
	0	0	0	0	1	0	N
	0	0	0	0	1	0	7
	0	0	0	0	0	1	8
	0	0	1	0	0	0	N
	0	0	0	0	1	0	9
	0	0	1	0	0	0	10
point B:	0	0	0	0	1	0	N
	0	0	0	1	0	0	N
	1	0	0	0	0	0	N
	0	0	0	0	1	0	1
	0	0	0	0	0	1	N
	0	0	0	0	0	1	N
	0	0	1	0	0	0	N
	0	1	0	0	0	0	N
	0	0	1	0	0	0	2
	0	0	0	1	0	0	N
	0	0	0	0	1	0	3
	0	0	0	0	0	1	N
	0	0	0	0	1	0	Q
	0	0	0	1	0	0	N
	0	0	0	0	1	0	Q
	0	0	0	0	1	0	Q
	0	0	0	1	1	0	4
	0	0	0	0	1	0	Q
	0	0	0	0	0	1	5
	0	0	1	0	0	0	N
	1	0	0	0	0	0	N
	0	0	0	0	1	0	Q
	0	0	1	0	0	0	6
	0	1	1	0	0	0	7

Quota matrix: 1 = (age/gender) young/male; 2 = young/female; 3 = middle/male; 4 = middle/female; 5 = old/male; 6 = old/female

N = No interview because not at home or not willing

Q = No interview because quota is saturated

In sampling point A the given quota characteristic had no real influence on the process of interviewing. In 18 households all persons living in these households were listed. In 10 households an interview was executed. Persons with characteristics 3 and 5 are contacted most frequently in this sampling point. But the concentration of reachable persons with the characteristics 3 and 5 did not influenced the process of interviewing: No target person was lost because a fulfilled quota matrix forbade an interview. In sampling point A the length of the random walk is induced by refusals.

In contrast, in sampling point B losses of interviews are induced by fulfilled quota matrix. Altogether 24 households are listed but only 7 interviews could be done. Persons with the characteristic 5 who were at home and were willing to give an interview were found by the interviewers very often. Without quota restrictions persons containing characteristic 5 could be interviewed four or five times as could be seen in Table 3: The interviewers were instructed by quota matrix to refuse interviews with persons containing the characteristic 5 after they had interviewed the second person with this characteristic. That happened by the third realized interview. Later on at their walk the interviewers were looking for young or middle aged persons but they met characteristic - 5 - persons, old men, who were bored sitting at home and willing to be interviewed. The list of the contacted persons in sampling point B shows the problematic of an uncontrolled random walk with a net number of interviews. But the walk controlled by a quota matrix is not the best solution. If all contacted households were listed - but they were not - then the quota matrix in the case of point B represents an exact reproduction of reality. There, one person is listed with characteristic 1 and one person with characteristic 2 and in each of the other four characteristics two persons are listed. Thus the quota matrix is a good instrument to control the walk of the interviewers.

However only two characters of the target persons were listed: age and gender. And the most important character is "just at home". This character is not listed because this character generates a bias. Most of the employed people are not at home at a normal working day around noon. But the interviewers doing quota-interviews are normal working people, starting work at nine in the morning and finishing daily work at five or six o'clock in the evening. If the interviewers were forced to contact the target persons of the households number 1 to 10 four times at the minimum at different time of day, even in the evening, before they were allowed to move an address to the category of the losses, then also other types of persons were interviewed. But interviewers who do quota interviews contact a household only once.

6 Analysis

A comparison of the three different sampling designs (see Table 1) shows:
The more the sampling design restricts the interviewer in doing the random walk, listing households and selecting persons for interviewing, the better the distribution of the sampled persons reflects the distribution in the population. The controlled sampling design contains the strongest restrictions: the listing of the addresses is selected from the selection of the target persons. Both steps can be controlled. The number of addresses who has to be listed is fixed by specification. A maximum of the target persons has to be interviewed. The uncontrolled random route is relaying on the interviewer's honesty and quality. Nearly nothing can be controlled, neither the random walk and the listing of the addresses nor the selection of the target persons. Normally an interviewer will contact a household only once and he or she will interview those persons who are at home and who are willing to answer the interviewer's questions. In this sampling design the interviewer passes by a lot of households to find persons for interviews. The quotation is a cosmetically restriction of weighting by age and gender.

In Table 4 the distribution of the interviewes by age, gender, and education is shown for each of the three different sampling designs:

Column 1 shows results from the German Micro Census as reference. The Micro Census is an annual survey conducted by the Federal Bureau of Statistics and the Statistical Offices of the Federal States of Germany with a sample size of one percent of the population living in Germany. Participation is obligatory for all persons who are selected in the sample. Therefore, the Micro Census nearly shows the real distribution of demographic characteristics of the whole the population.

Column 2 shows the distribution of demographic characteristics produced by the controlled sampling design, column 3 produced by the uncontrolled design and column 4 that produced by the random route plus quota design.

Before discussing the distributions, a short comment about the used data-sets is necessary.

The controlled and the uncontrolled design are represented by data-sets from the German General Social Survey (ALLBUS) from 1998 (controlled) and 1992 (uncontrolled). The definition of the population is: persons in the age of 18 years and older living in private households. The random route plus quota design was used in a national survey of the German Youth Institute. The population is defined as persons aged between 18 and 55 years, living in private households. Therefore the age groups as well as the education groups in column 4 are not really comparable with the distributions in the other columns.

Table 4: Distribution of demographic variables in different sample designs: West-Germany population of 18 years and older.

	Micro-Census	controlled design	uncontrolled design	RR plus Q design
Gender				
male	47.5	49.6	49.0	44.7
female	52.5	50.4	51.0	55.3
N	48,362,000	4,051	4,680	8,790
Age				
18 – 29	17.1	17.0	21.7	23.9
30 – 39	19.7	19.8	21.3	30.5
40 – 49	16.8	16.7	17.5	29.5
50 – 59	17.0	18.8	17.6	16.1
60 – 69	14.7	17.0	13.1	
70 +	14.6	10.7	8.8	
N	48,362.000	4,048	4,680	7,972
Education				
9 years	56.8	49.9	53.1	43.8
10 years	23.3	28.3	25.2	34.2
12/13 years	19.9	21.8	21.7	22.0
N	44,575,000	3,851	4,451	8,567
HH-Size				
1-Pers.-HH	20.7	15.2	12.4	17.2

Allbus 1992, 1998; MZ 1997; FamilySurvey 2001

Comparing the distributions shown in columns 1 to 3 shows that with respect to “age” the data-distribution of the controlled design are closer to the distribution of Micro Census data than the data-distribution of the uncontrolled design. With respect to education, the controlled design shows a worse data-distribution than the uncontrolled design. The controlled design produces an over-sampling of the middle educational level because those members of the old age groups who will respond are more often the higher educated persons of that groups. In the uncontrolled design the older respondents are also higher educated but the easy to reach persons of the younger age groups are more often lower educated.

7 A comparison of pairs

In a second step only those persons were part of the analysis who were living in a two persons household (in the age of 18 years and older) together with a partner of the opposite gender. In this type of household the survey population is represented in each household exactly with one male and one female person. Children in the age below of 18 are out of interest for this analysis.

In a random sample of pairs there should be a rectangular distribution of male and female persons. And with the distribution of age a likewise distribution of male and female can be anticipated, even if pairs often are not exact in the same age.

Table 5: Demographic distribution of respondents in households of pairs: Exact one man and one woman in the age of 18 years and older.

	Controlled Design		Uncontrolled Design		RR plus Q Design	
	Percent	Deviation	Percent	Deviation	Percent	Deviation
Pairs without Child						
male	54.0	+4.0	56.4	+6.4	47.0	-3.0
Female	46.0	-4.0	43.6	-6.4	53.0	+3.0
N	739		870		1,920	
Pairs with Child						
male	46.3	-3.7	47.0	-3.0	37.7	-12.3
Female	53.7	+3.7	53.0	+3.0	62.3	+12.3
N	739		940		2,045	
Female by Gender and Age						
20 – 29	61.9	+11.9	63.8	+13.8	68.1	+18.1
30 – 39	59.0	+9.0	58.7	+8.7	61.4	+11.4
40 – 49	46.3	-3.7	56.2	+6.2	51.0	+1.0
50 – 59/55	49.1	-0.9	44.8	-5.2	52.7	+2.7
60 – 69	40.5	-9.5	31.6	-18.4		
70 +	37.4	-12.6	26.7	-23.3		

Sources: Allbus 1992, 1998; FamilySurvey 2001; data unweighted

The sampling design has a significant effect as can be seen in Table 5. The gender distribution can be corrected by quota. But as can be seen in Table 5 in the group of pairs with a child the easy to reach are the women. Here the interviewer doing quota design is forced to contact one household only once. If an interviewer is starting at 9 o'clock in the morning doing interviews for 8 hours he or she must reach more women with children then without.

The last part of Table 5 shows the easy to reach groups. In the younger cohorts these are the female and in the older cohorts these are the male persons. The more uncontrolled a sampling design is the more female persons in the younger cohorts were interviewed. The quota cannot control this trend very much. In the oldest two cohorts women are under-represented, even those who are not living alone.

8 Conclusion

The hypothesis is confirmed by the data:
If the sampling design offers a wide scope for an interviewer's uncontrolled selection of the target persons then the groups of easy to reach persons are over-represented. In these cases there is a large distance between the distribution of survey data and the real distribution in the population. The new sampling design combining a random walk with quota cannot correct the failures given by contact only easy to reach persons - in the opposite: a combination with quota allows the interviewer to forget all instructions about contacting also difficult to reach persons. The interviewer will no longer contact a household twice. The walk gets longer and longer in the extreme case the interviewer has to contact about 100 households for realizing 10 interviews. But the unproductive and bad paid time of long ways, e.g., for coming back for only one missing interview has decreased very much. Now, the interviewer will not work any longer at evenings and weekends. Random route combined with quota allows a full-time interviewing person to do a normal job on five days a week from 9 a.m. to 5 or 6 p.m.

The problem is that difficult to reach persons are other type of persons than easy to reach persons. E.g., young mothers in their role as housewives have other themes to speak about or other attitudes than young ladies struggling for their occupational career. And unemployed men, who are reachable at home during daytime shows different attitudes or behavior then men from the same cohort who are employed. From the Micro Census we only get a regional distribution of age and gender for controlling our survey samples. Therefore at the end we get biased data by survey research and only the knowledge of the sampling design what was practiced can help to classify the quality of the data.

References

[1] Alreck, P.L. and Settle, R.B. (1995): *The Survey Research Handbook*. 2nd ed. Chicago: Irwin.

[2] Arbeitsgemeinschaft ADM-Stichproben und Bureau Wendt (1994): Das ADM-Stichprobensystem Stand: 1993; In S. Gabler, J.H.P. Hoffmeyer-Zlotnik, and D. Krebs (Eds.): *Gewichtung in der Umfragepraxis*. Opladen: Westdeutscher Verlag: 188-202.

[3] Fink, A. (1996): *How to Sample in Surveys*. 2. print.. Thousand Oaks, Calif.: Sage.

[4] Foreman, E.K. (1991): *Survey Sampling Principles*. New York: Dekker.

[5] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953): *Sample Survey Method and Theory*, 1/2. New York: Wiley.

[6] Hartmann, P. (1990): Wie repräsentativ sind Bevölkerungsumfragen? Ein Vergleich des ALLBUS und des Mikrozensus. *ZUMA-Nachrichten,* **26**, 7-30.

[7] Hoffmeyer-Zlotnik, J.H.P. (1997): Random-Route-Stichproben nach ADM; In S. Gabler and J.H.P. Hoffmeyer-Zlotnik (Eds.): *Stichproben in der Umfragepraxis.* Opladen: Westdeutscher Verlag, 33-42.

[8] Kirschner, H.-P. (1984): ALLBUS 1980: Stichprobenplan und Gewichtung. In: K.U. Mayer and P. Schmidt (Eds.): *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften.* Frankfurt a.M./New York: Campus, 114-182.

[9] Kish, L. (1949): A Procedure for Objective Respondent Selection wtithin the Household; *Journal of the American Statistical Association*, **44**, 380-387.

[10] Kish, L. (1965): *Survey Sampling.* New York: Wiley.

[11] Levy, P.S. and Lemeshow, S. (1991): *Sampling of Populations: Methods and Applications.* New York: Wiley.

[12] Raj, D. (1968): *Sampling Theory.* New York: McGraw-Hill.

[13] Rösch, G. (1994): Kriterien der Gewichtung einer nationalen Bevölkerungsstichprobe. In: S. Gabler, J.H.P. Hoffmeyer-Zlotnik, and D. Krebs (Eds.): *Gewichtung in der Umfragepraxis*. Opladen: Westdeutscher Verlag: 7-26. Schaefer, F., 1979: Muster-Stichproben-Pläne. München: Moderne Industrie.

[14] Schmidtchen, G. (1961): *Die repräsentative Quotenauswahl: Bericht über ein Quota-Random-Experiment des Instituts für Demoskopie Allensbach.* Allensbach : Inst. f. Demoskopie.

[15] Som, R.K. (1966): *Practical Sampling Techniques.* 2. ed., rev. and expanded. New York: Dekker.

[16] Sudman, S. (1976): *Applied Sampling.* New York: Academic Press.

[17] Tryfos, P. (1996): *Sampling Methods for Applied Research: Text and Cases.* New York: Wiley.

[18] Wright, S.R. (1979): *Quantitative Methods and Statistics: A Guide to Social Research.* Sage.