

POSDEM: FRAME AND RELIABILITY

Gonzalo Sánchez-Crespo Benítez

Instituto Nacional de Estadística, Madrid, España; gsanchez@ine-es

Abstract

Try to reduce the sampling error for a specific survey population frame with the stratification and the sample size done is possible at any time. The sampling methods are tested over the frame with the purpose to make that the sample will be more representative. POSDEM allows testing 28 different sampling methods over any population frame. Look for increasing the sample heterogeneity. Software is open and can be added more sampling procedures. It can be used in surveys, censuses, quality control or to explore frames into big data.

Sampling methods have a strong relationship with the survey population frame. Also there are many sampling methods not frequently apply. For different samples size, software POSDEM, test these relationship between frames and sampling methods. That test might be done just over the real population or over a super population model inspired on this population. We find interesting results concerning procedures stability properties.

We know that several alternatives of systematic sampling and unequal probabilities sampling are frequently used. However both may be disturbed by the structure of the population investigated. Trends, cycles and inappropriate relationship between variables are present more often that it seems. In both cases, to find the best sampling method, it is necessary to study the relationship between the sampling plans and the population frame before applying these methods.

Keywords: Systematic sampling. Probabilities proportional to size. Superpopulation model.

1. Introduction

POSDEM, allows for 28 sampling methods, answering among other questions: what is the best sampling plan for a specific population frame?; what is the sampling size related for a significance level?; what is the mean square error expected?; what are the units that must be investigated?; what happens if ...? (Biehler, R.).

The results obtained with the various available procedures of systematic sampling could depend on several factors such as the structure of the population and the hypothesis made (Bellhouse, D.R. and Rao, J.N.K.).

An example has been prepared in order to show the possibilities of this software. This paper examines several sampling methods to know what happens if in the structure of the population unexpected changes occur. For instance, it may be expected that population presents a linear trend but it actually has a polynomial trend, because there has been changes in the structure of the population since the last available information. Is there any sampling method robust enough when the population hypothesis made to apply the sampling methods have unexpected changes? With this software, this question can be answered. Another interesting feature is the behavior of the sampling methods when the degree of randomness changes.

Considering the observed erratic behavior, attention has been focused on the centered located systematic method, when the sample size increases as well as on the inconvenient need to differentiate if the sampling interval, $k=N/n$, is even or odd. Until now, only the Yates method of extreme corrections eliminated the linear trend for even or odd values of the sample size. The centered method of Madow does not eliminate the linear trend when the sampling interval, k , is even, and the balanced and modified systematic sampling procedures do not eliminate the linear trend when the number of units in the sample is odd.

2. Population and methods considered

The population presented by Murthy, M.N. (1967, p.127) and used by Krishnaiah, P.R. and Rao, C.R. (1988, p.131, population 4), will be applicable to illustrate the changes due to the fit of polynomial models with degrees one to five. The purpose is to check how the changes in the specification of the model, in the sample size or in the distribution of the random error, affect the mean squared error. The population considered consists of 128 units from the 1961 census.

The following methods have been analysed: random stratified sampling with one sample unit per stratum(0), systematic sampling with constant sampling interval (1), systematic sampling corrected in the extremes of Yates (2), balanced systematic sampling (3), modified systematic sampling (4), and centered systematic sampling with the constant interval of Madow (5) (). To these classical methods we add two new methods: systematic sampling with a variable sampling interval and its application to the centered method. This method will be referred to as

centered systematic sampling with a variable interval (6). We have been respectful to the numbers that Bellhouse and Rao gave to the methods.

Other methods, like probabilities proportional to size sampling methods (Sánchez-Crespo Rodriguez, J.L.), considered by POSDEM but not directly related to this paper, could be found in the following internet page: <http://goo.gl/B3CC83>

With POSDEM it is possible to connect these methods between them. Four basic methods have been considered: constant interval, variable, balanced and modified. And they have been to combine with the centered method and with the extreme corrections making a total of sixteen systematic methods.

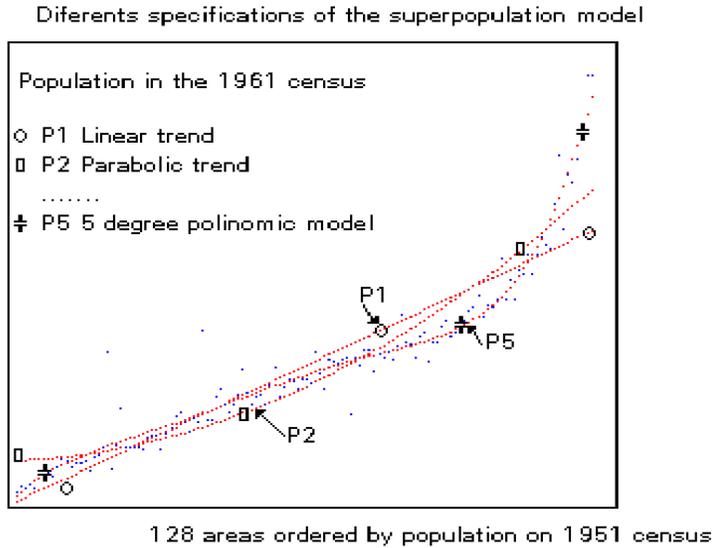
The comparison of the systematic centered with a variable sampling interval, involving the methods of unequal probability sampling, will be left for a second article.

3. The superpopulation models

With the population considered the best fit corresponds to the polynomial model with five degrees (P5). Nevertheless it is also possible to suppose that the P1 and the P2 models represent the population with the purpose of seeing what would happen if the hypothesis that we made about the population changes.

With these models we are able to make a set of random populations that are in agreement with a patron. In each of these populations so generated it is possible to obtain all the possible samples and determine the mean square error of the estimator for different selection procedures. The set of the mean square error so calculated allows us to compute the expected value of the mean square error and its variance over the model.

Fig.1: Different models for the same population



4. Evaluation of sampling methods

To evaluate the sampling methods it is applied a superpopulation model adapted to each population. Some of the methods considered, centered and ends corrections, are biased so it is necessary to determine the mean squared error (mse).

$$mse(\hat{x}) = E \left(\frac{\hat{x} - X}{x - X} \right)^2 \quad (1)$$

POSDEM get the structure of the population fitting the data with the orthogonal polynomial method. Next the software generates populations defined by the model calculating the mse for each population and sampling method considered. The mean of the mse so calculated, over the set of populations generated, will be the approach to the expected value of the mse under the superpopulation model.

$$E^*(mse(\hat{x})) \cong \frac{\sum_{g=1}^G mse(\hat{x})_g}{G} \quad (2)$$

Where $g=1,2,\dots,G$ represents the set of finite populations generated with the model.

This evaluation procedure assures the link of this work with the work of Bellhouse, D.R. y Rao, J.N.K. The results obtained for models of degrees one and two are coincident and let us suppose that for higher degrees it will also be coincident. The symbols used in the results tables will be $E^*(mse(\text{number}))$, where E^* is the expected value over the model and number represents the sampling method considered.

To know in which way POSDEM allows evaluate sampling methods it is necessary to see the next part call mean squared error stability: A new upper confidence limit to evaluate sampling methods.

5. Mean squared error stability: upper confidence limit to evaluate sampling methods.

The variance over the model is considered the measure of the accuracy for the mean squared error. And it is done by the following expression:

$$V^*(mse(\hat{x})) \cong \frac{\sum_{g=1}^G \left(mse(\hat{x})_g - E^*(mse(\hat{x})) \right)^2}{G} \quad (3)$$

It could be observed that if the random error of the model increases then the centered method would become erratic, because the variance over the model increases too. This is the reason to evaluate the sampling methods using an upper confidence limit trough the expected mse value and his deviation . The expression results $+1.96$ allows to have a confidence limit about that the expected error will not be greater than that this upper confidence limit, at least in 95% of 100 populations generated with this model. The sampling method will be selected considering the expected value of the mse and his variance over the superpopulation model.

6. The original population and its relation to the superpopulation approach

This table is introduced with the aim of showing that the calculations carried out by the computer program POSDEM are done correctly.

Table 1: The mean square error for different methods in the original population presented by Krishnaiah, P.R. and Rao, C.R. (1988, p.131, population 4)

	n=4	n=8	n=16	n=32
Systematic constant interval (1)	487,706.63	133,189.25	39,648.80	5,884.60
Centered with const.inter.	66,012.53	2,330.25	8,664.66	94.76
Extreme corrections	69,135.45	19,228.90	3,401.43	1,152.19

Table 2: Shows the relationship between the mean square error computed in the natural population, the expected value taken over the model and the variance of the mean square error over the model. If a selection procedure is chosen regarding the result obtained in columns (1), (2) or (4) we have columns (5) to (7) that show the preferable method for each hypothesis. The column (4) is a confidence limit of the mean square error over the model.

	(1)	(2)	(3)	(4)	Selection procedures choose using:		
$\sigma_e = 200$	mse	E*(mse)	V*(mse)	CL*=E*(mse)+2* $\sqrt{V*(mse)}$	mse	E*	CL*
Cwci_n=4	66,012.53	21,624.19	313,842,251.65	57,055.38	Cwci	Cwci	Cwci
Ec_n=4	69,135.45	47,383.00	42,475,069.38	60,417.58			
Cwci_n=8	2,330.25	6,781.00	38,985,713.00	19,268.71	Cwci	Cwci	
Ec_n=8	19,228.90	11,213.00	8,841,197.00	17,159.83			Ec
Cwci_n=16	8,664.66	3,807.60	11,919,358.68	10,712.48			
Ec_n=16	3,401.43	2,812.13	2,145,672.35	5,741.75	Ec	Ec	Ec
Cwci_n=32	94.76	2,232.56	4,461,746.87	6,457.13	Cwci		
Ec_n=32	1,152.19	969.48	571,906.83	2,481.97		Ec	Ec

In the example where $n=8$ the expected value of Centered with Constant interval is clearly better than the Extreme Corrections. Nevertheless, the variance over the model is considerably different for these methods and when choosing them with the proposed Upper Confidence Limit the results become better for the Extreme Corrections.

7. Populations structures under the superpopulation models with POSDEM

Here the module of simulation of structures is used in the application POSDEM . This makes it possible to define a superpopulation model, which takes into consideration the population used, with the expression:

$$X_u = a_0 + a_1 U_1 + a_2 U_2 + a_3 U_3 + a_4 U_4 + a_5 U_5 + e_u \quad (4)$$

Where U represents the population units that in this case takes values between 1 and 128. a_i with $i=1, 2, 3, 4$ and 5 are the parameters computed by the least squared method. e_u is a random error term where $E^*(e_u)=0$; $E^*(e_u e_v)=0$ with $u \neq v$; $E^*(e_u^2) = \sigma_e^2$, E^* represents the expected values in respect to the model.

Fig.2: POSDEM screen to modelized original populations

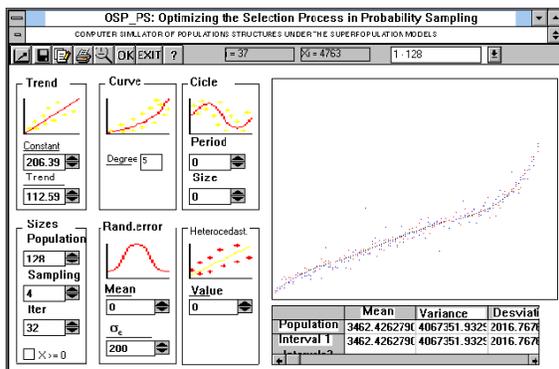


Table 3: The table below shows the estimators obtained for different degrees of the parabolic model used to represent the population showed by Murthy, M.N. (1967,p127).

Models	a_0	a_1	a_2	a_3	a_4	a_5

P1	149,08	51,38				
P2	1.179,16	-3,84	0,368			
P3	-83,03	119,02	-1,854	0,0115		
P4	728,48	-2,54	2,347	-0,0390	-0,0002	
P5	206,39	112,59	-3,803	0,0873	-0,0009	-0,000003

The results obtained when this model, using the random generation of populations, must be in agreement with the following theoretical results provides by Bellhouse, D.R. and Rao, J.N.K:

A) For the linear case:

1) Systematic sampling: $E^* V_p(\hat{x}_{sis}) = a_1^2 (k^2 - 1)/12 + \bar{\sigma}^2$ with $\bar{\sigma}^2 = \sigma_e^2(k - 1)/nk$

The first variance component of the estimator is due to the linear trend, and the second to the random error.

2) Random sampling: $E^* V_p(\hat{x}_{sr}) = a_1^2 (k - 1) (nk + 1)/12 + \bar{\sigma}^2$

3) Stratified sampling with one unit for stratum: a population with n strata formed by the next sets, is used: $\{1...k\} \{k+1,...2k\} \dots \{(n-1)k+1, \dots, nk\}$. $E^* V_p(\hat{x}_{str}) = a_1^2 (k^2 - 1)/12n + \bar{\sigma}^2$

It is easy to see that in this case: $E^* V_p(\hat{x}_{strat}) \leq E^* V_p(\hat{x}_{sis}) \leq E^* V_p(\hat{x}_{sr})$

B) For the non linear case:

The following formula is used to check results when the degree of the polynomial model is two. $E^*(mse(2)) - E^*(mse(5)) = (a_2^2/720)(k^2 - 1)(19k^2 - 31) > 0$ for k odd

The mean square error is used because the centered (method 5) and corrected (method 2) procedures are biased methods.

In conclusion, with small samples (5) is better than (2) and this trend is reversed for large samples. When there is doubt between the application of (5) or (2), method (6) should be used since it is more stable than (5) and (2) in relation to changes in the sampling size.

8. Conclusions

The main results of this work are:

1.- The implementation of a computer program to carry out the optimisation of the sampling selection and to evaluate alternative designs for specific populations under the superpopulation approach. This program could be useful for students studying the theory and practice of sampling in finite populations, in both their basic and advanced versions; for professors that want to provide an instrument for the empirical investigation over the area of sampling surveys; and for central statistical offices, companies, or researchers. This new program is currently available for use on the Internet. <http://goo.gl/LjHGm1>

2.- The results obtained using these sampling methods on various populations, are in agreement with the theory. However, we found an interesting result: sampling methods, which are excellent in some situations, are very bad in others. The new systematic sampling method proposed, centered with variable interval, behaves well in different situations. In various examples, it reduces the mean square error by comparison to other systematic procedures considered. This reduction is due strongly to the variations of the sampling size, the numbers of groups in the population, or if they are even or odd. This method also works when, in the population, a structure change occurs. This avoids the risk of choosing a sampling method which behaves badly if the population has unexpected changes.

3.- In view of the experimental increases in the random nature of the population, the centered methods presents a small expected error, as the theory says, but with a large variability, as practice shows. This suggests a new indicator to measure the sampling error to be used to improve the evaluation of systematic sampling methods. This indicator is an upper confidence limit of the mean square error distribution over the superpopulation model. So it is possible to

explain and control the erratic behavior observed in the centered systematic sampling in relation to the random term of the considered model.

Reduce the sampling error for a specific survey population frame with the stratification and the sample size done it is possible if we increase the sample heterogeneity. Sampling methods have a strong relationship with the survey population frame. POSDEM test sampling and frames to find best sample that mean less error limits.

9. References

Bellhouse, D.R. & Rao, J.N.K. (1975); Systematic sampling in the presence of a trend; *Biometrika* 62. pp.694-697;

Biehler, R. (1997). Software for learning and doing statistics. *International Statistics Review*, pp 167-189

Murthy, M.N. (1967): «Sampling theory and Methods» Statistical Publishing Society, Calcutta. pp.127

Sánchez-Crespo Benítez, G (1998); Muestreo sistemático con intervalo variable; *Estadística Española* 143. pp.2-32; <http://goo.gl/E44Bxv>

Sánchez-Crespo Benítez, G (1999); POSDEM; *Revista electrónica de Metodología Aplicada*, Vol 4 n 2, pp 12 -36; <http://goo.gl/SkIYq6>

Sánchez-Crespo Rodríguez, J.L. (1997). A Sampling Scheme With Partial Replacement. *Journal Official Statistics*, 13, 4, pp 327-339. <http://goo.gl/JGHKvB>

Krishnaiah, P.R. & Rao, C.R. (1988): «Sampling», *Handbook of Statistics*, North-Holland. pp.131