Pakistan Journal of Applied Sciences 2(1): 36-38, 2002

Sampling Designs Based on Subpopulations

Farmakis Nicolas
Aristotle University of Thessaloniki, Department of Mathematics
GR-54006 Thessaloniki, Greece

Abstract: Some low budget designs for sampling in m-dimensional subsets (m=1,2,3,...,k-1) of some k-dimensional populations are proposed. Elements from the selected subset of the population (instead of the whole population) are drawn by one of the classical and well known sampling techniques, like the Simple Random Sampling or the Systematic one, etc. The final budget comes down over 10 times. Some illustrative examples are given.

Keywords: Sampling, Random, Systematic, Population, Designs, Multi-dimensional

Introduction

In the case of a Population **P** with elements (units) taken as points of a k-dimensional Domain **D**, i.e. $\mathbf{D}\subseteq R^k$, the need of a low budget sampling arises. Every unit $\mathbf{u}\in\mathbf{D}$ figures as the $\mathbf{u}(x_1,x_2,...,x_k)\in\mathbf{D}\subseteq R^k$, where $x_j\in R$ is the j-th coordinate of **u**, j=1,2,..., k. We study the random variable Z, defined in **D**, as

$$Z(\mathbf{u})=\mathbf{f}(x_1,x_2,...,x_k)\in R,\ \mathbf{u}(x_1,x_2,...,x_k)\in\mathbf{D}. \tag{1.1}$$

Obviously we work on the (k+1)-dimensional Euclidean Space, the R^{k+1}, and deal with a graph (Z) related to the k-variable function **f**. We assume that we have some knowledge about the form of $\mathbf{f}(x_1,x_2,...,x_k)$ subject -probably- to some parameters. It is well known that the mean value of $Z(\mathbf{u})$ is given by

$$\mu_z=\overline{Z}=\frac{1}{\|D\|}\iint_D\!\!\cdots\!\int f(x_1,x_2,x_3,...,x_k)\,dx_1dx_2dx_3...dx_k, \tag{1.2}$$

where $\|D\|$ is the value of the Euclidean measure (hypervolume) of **D**. All we propose in this paper can be used (at least) for the purposes of the administration in the branches of forestry, fishery, agriculture, environment, etc., etc.

The main goal of this paper is the estimation of μ_z, based only on a sample of points (units of **P**) taken from an arc $C\subseteq\mathbf{D}$ with dimension m, $1\le m<k$, instead of points taken from whole the domain **D**. This can make the budget many-many times (more than 80 probably) lower than the budget in the classical sampling from the whole domain **D**. Note that in almost all the cases of these sample designs m is equal to 1.

The Sampling Design: The new idea in this paper is the use of analytical methods as tools for the identification of a suitable subpopulation of dimension m<k, which will serve as the population for sampling. This subpopulation is a subset of a subspace (C) of R^k. We are going to draw elements only from the arc C=(C)∪**D**. If the form of **f** in (1.1) involves unknown parameters, then a presampling procedure may be used to provide point or (better) interval estimators for the unknown parameters. In any case we need a good deal of information about the shape (when k=2) or, more generally, about the form of (1.1) from previous surveys or from any other source, like presampling procedures, archives, etc.

Naturally, the best scenario is when the exact value of the integral in (1.2) is known. If the integral does not exist or it cannot be calculated by a finite number of analytical operations, then we need the help of some numerical methods. In Farmakis, (2001) and Farmakis, (1999) the basic steps of the method we propose to designate the arc C=(C)∪**D**, for the estimation of $\mu_z=\overline{Z}$, are given. In general lines these steps are:

STEP 1: Collect as much information as possible about function $Z(\mathbf{u})=\mathbf{f}(x_1,x_2,...,x_k)/\mathbf{u}\in\mathbf{D}$. Also figure it in an analytical way as accurate as it is possible (make a regression analysis or use other approximation methods to do it). Use the more simple case of function you trust on.

STEP 2: Evaluate $\mu_z=\overline{Z}$ given in (1.2) by integration or approximation.

STEP 3: Consider the arc C=(C)∪**D**. The curve (C) is a m-dimensional subspace of R^k, m<k, i.e. C is the set of points (elements):

$$C=\left\{(x_1,x_2,...,x_k)\,x_{r_{m+j}}=\varphi_{r_{m+j}}(x_{r_1},x_{r_2},...,x_{r_m})\,x_{r_i}\in I_{r_i}\subseteq R,\ i=1,...,m;\ j=1,...,k-m\right\} \tag{2.1}$$

and denote the measure (norm) of C, as $\|C\|\in R$. Observe that arc $C\subseteq R^m\subseteq R^k$.

STEP 4: Evaluate the mean value of Z over C, namely

$$\overline{Z}_{(C)}=\frac{1}{\|C\|}\iint_{(C)}\!\!\cdots\!\int f\left(x_{r_1},x_{r_2},...,x_{r_m},\varphi_{r_{m+1}}(x_{r_1},...,x_{r_m}),...,\varphi_{r_k}(x_{r_1},...,x_{r_m})\right)dx_{r_1}dx_{r_2}...dx_{r_m} \tag{2.2}$$

STEP 5: Look at the difference $\Delta\overline{Z}=\left|\overline{Z}_{(C)}-\overline{Z}\right|$ and solve one (the suitable in any case) of equations (2.3), in order to obtain the formula of the curve (C):

$$\Delta\overline{Z}=0 \tag{2.3}$$

$$\Delta\overline{Z}=\Delta_0\overline{Z}=\min_{(C)}\Delta\overline{Z} \tag{2.3a}$$

$$\frac{\Delta\overline{Z}}{\overline{Z}}\le d,\ d\in(0.00,a),\ a\le 0.20 \tag{2.3b}$$

STEP 6: Take a sample of points (e.g. systematic) on

the arc C and calculate the estimator $\hat{\bar{Z}}'$ of $\bar{Z}=\mu_z$ (Cochran, (1977), Farmakis, (1999 and 2000).

Applications: In several cases the traditional sampling techniques are both time and money consuming. The over proposed technique which does not suffer from the above problems can be used, among other settings, in agriculture where the sampling from an entire field can be replaced by the sampling on a single straight path. Also we can refer to many uses in Biology, Environmental research, Economy, etc. These comments as well as the next examples show that the contribution of the present work lies on the fact that the proposed technique is extremely useful due to both its simplicity and its low budget.

Some examples will go to illustrate the usefulness of the above-proposed method of ours. They are based on Simple random and on systematic sampling techniques as in Cochran, (1977), and Farmakis, (2000).

Example 3.1: Take k=2, m=1, Z=f(x,y)=axy, and D={(x,y), $0\le x,y\le t$ }.

Answer: The proposed method gives, as a solution of $\Delta\bar{Z}=0$, the line (ε) y=0.75x, $0\le x\le t$. The suitable arc (space) of this line is C={(x, 0.75x), $0\le x\le t$ }, with $\bar{Z}=0.25\cdot a\cdot t^2$. A random sample of 32 values of x, obtained by the CASIO fx-85VH pocket calculator, is:

S={0.087t, 0.323t, 0.733t, 0.743t, 0.708t, 0.061t, 0.500t, 0.462t, 0.595t, 0.908t, 0.090t, 0.159t, 0.439t, 0.928t, 0.071t, 0.382t, 0.453t, 0.338t, 0.243t, 0.598t, 0.136t, 0.015t, 0.871t, 0.191t, 0.265t, 0.218t, 0.343t, 0.913t, 0.240t, 0.479t, 0.069t,0.822t}.

Simple calculations show that $\hat{\mu}_{z,RS}=\hat{\bar{Z}}=0.2554002\cdot a\cdot t^2$.

Example 3.2: In the case described in Example 3.1 take a systematic sample of 32 points.

Answer: The points are $S_{sy}=\{x_r,\ x_r=\frac{r\cdot t}{31},\ r=0;1,2,...,31\}$ and of course it is $z_r=0.75\cdot a\cdot t^2\cdot\frac{r^2}{31^2}$.

So we obtain $\hat{\mu}_{z,Sy}=\hat{\bar{Z}}=0.2540322\cdot a\cdot t^2$.

Example 3.3: Take a systematic sample of size n in the case described in examples 3.1.

Table: 3

t ↓	r→	1μ	6μ	9μ	12μ
2 sec		9480900	9141200	8943100	8749400
3 sec		9265400	8933300	8740000	8550000
5 sec		8849000	8532100	8346500	8116000

Answer: (a) From (3.1) and (3.2) we get the general function Z(t,r) giving the number of the microbes per μ^2 related to the time t and to the distance r from the origin O(0,0):

$Z(t,r)=m_t(r)=M_t\cdot e^{-ar}=M_0\cdot e^{-bt}\cdot e^{-ar}$

or

$m_t(r)=Z(t,r)=M_0\cdot e^{-(ar+bt)}$ (3.3)

Answer: The points are $S_{sy}=\{x_r,\ x_r=\frac{r\cdot t}{n-1},\ r=0,1,2,...,n-1\}$ and of course it is $z_r=0.75\cdot a\cdot t^2\cdot\frac{r^2}{(n-1)^2}$. So we obtain $\hat{\mu}_{z,Sy}=\hat{\bar{Z}}=0.25\cdot a\cdot t^2\cdot\frac{2n-1}{2(n-1)}$. A simple view on the next relation gives a very interesting result:

$\hat{\mu}_{z,Sy}\xrightarrow[n\to\infty]{}0.25\cdot a\cdot t^2=\mu_z$.

A more useful for theoretical reasons case will be the next one in example 3.4:

Example 3.4: Take $Z(\mathbf{u})=\mathbf{f}(x,y)=\left(a^2-x^2-y^2\right)^{0.5}$, with $\mathbf{D}=\{(x,y),x^2+y^2\le a^2\}$. Give an estimation for $\mu_z=\frac{2}{3}\cdot a$ (Farmakis, 1999.)

Answer: After some calculation we obtain by the above proposed method as a solution the curve :

(C) : $y=0.5287\cdot a$ and $(x,y)\in\mathbf{D}$.

A systematic sample of n=101 points gives an estimation of μ_z:

$\hat{\mu}_z(101)=0.659368\cdot a$.

Also for n=201 points it is:

$\hat{\mu}_z(201)=0.663101\cdot a\cong\frac{2}{3}\cdot a=\mu_z$.

The next example shows how we can manipulate some presampling data, if we know something about the nature of them, e.g. an exponetial model connecting the variables. The data are imaginary.

Example 3.5: The number of the microbes crowed, at the moment t, per every area $ds=1\mu^2$, is a function of the distance $r=\left(x^2+y^2\right)^{0.5}$ from the origin (0,0). This number is given by

$m_t(r)=M_t\cdot e^{-ar}$, $0\le r=\left(x^2+y^2\right)^{0.5}\le 20$, $0\le t\le 10$, $M_t=m_t(0)$. (3.1)

Also the number of the microbes at the origin at the moment t (sec) is

$M_t=m_t(0)=M_0\cdot e^{-bt}$, $0\le t\le 10$. (3.2)

We have got the results of n=12 observations (Table 3.1).

(a) Find an estimation of M_0, a and b in (3.1) and (3.2).

(b) Propose a generalization of the result and use it to make a sampling design in order to save money and keep the variance of the sample mean $\bar{m}_t$ as low as the situation allows.

We are going to try regression on the data of table 3.1: From (3.3) we get

$\ln Z(t,r)=\ln M_0-a\cdot r-b\cdot t$ (3.4)

After some calculations the results of regression analysis are :

$$\begin{bmatrix} \ln \hat{M}_0 \\ \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 16.12051 \\ 0.00746 \\ 0.02355 \end{bmatrix}$$

i.e. $m_t(r)$ in (3.3) is estimated as

$$m_t(r) = Z(t,r) = 1.0024 \cdot 10^7 \cdot e^{-(0.00746 \cdot r + 0.02355 \cdot t)} \quad (3.5)$$

(b) The mean value of the number of microbes per μ² in the points of **D** is now estimated by

$$\hat{\bar{Z}} = \hat{\bar{m}} = \frac{1.0024 \cdot 10^7}{\|D\|} \iint_D e^{-(0.00746 \cdot r + 0.02355 \cdot t)} dr \cdot dt$$

=8297180 microbes/μ²

with **D**={(r,t): $0 \le r = (x^2 + y^2)^{0.5} \le 20$, $0 \le t \le 10$ }, $\|D\| = 200\mu \cdot \sec$. This is an estimation of $\bar{Z} = \bar{m}$ over all the point (elements) of **D** via regression and simple random sampling.

After this we are going to try for a straight line (ε) : $t = \lambda \cdot r$, $\lambda \in R$, in order to calculate the value of $\bar{Z}_{C'}(r) =$

$$\bar{m}_{\lambda \cdot r}(r) = \frac{1.0024 \cdot 10^7}{20} \cdot \int_0^{20} e^{-(0.00746 + 0.02355\lambda) \cdot r} dr,$$

$C = (\varepsilon) \cap D$ (3.6)

and to solve the equation (2.3), STEP 5, for the value of λ, i.e. $\bar{Z}_{C'}(r) = 8297180$, due to unbiased nature of simple random sampling estimators.

Equation (2.3) gives after some calculations as a solution the value λ=0.513 approximately, i.e. we have the line

(ε): $t = 0.513 \cdot r$ (3.7)

as the subpopulation of **D** which will give us the sample we need. From (3.6) and (3.7) we have the next form of estimation of $\bar{Z}$

$\bar{Z}_{C'}(r) =$

$$\bar{m}_{\lambda \cdot r}(r) = \frac{1.0024 \cdot 10^7}{20} \cdot \int_0^{20} e^{-0.01954 \cdot r} dr \quad (3.8)$$

as we face the case from the point of view of a continuous variable r.

We will use the set C= $(\varepsilon) \cap D$ as the population from which we will draw off a sample of size n=21 (systematic) to estimate the mean value of $m_t(r)$.

Thus we have:

$$\hat{\bar{m}} = \bar{m}_C = \bar{m}_{0.513 \cdot r}(r) = \frac{1.0024 \cdot 10^7}{21} \cdot \sum_{r=0}^{20} e^{-0.01954 \cdot r}$$

microbes/μ²

$$= 0.47733 \cdot 10^6 \cdot \sum_{r=0}^{20} e^{-0.01954 \cdot r} = 8302632 \text{ microbes/}\mu^2. \quad (3.9)$$

The above result is an estimation of $\bar{Z} = \bar{m}$. Since we have already found by the random sample and via regression the estimation $\hat{\bar{Z}} = 8297180$ microbes/μ², the new estimation in (3.9) seems to be a good one, with a relative error (difference) 0.0657% to $\hat{\bar{Z}} = 8297180$. We however note that the estimation in (3.9) is based only on n=21 points, saving time and effort (money).

Too many other cases like the present in this example can be manipulated by this method as it becomes obvious.

References

Cochran, W.G. 1977. Sampling Techniques, Wiley, New York.

Farmakis, N. 1999. Systematic Sampling in Two-dimensional Populations, Proc. of the 52nd International Congress of ISI, Contr. Papers B.-1, pp 319-320, Helsinki, Finland.

Farmakis, N. 2000. Introduction to Sampling. (in Greek) Christodoulidis Publications, Thessaloniki, Greece.

Farmakis, N. 2001. Sampling Designs in Multidimensional Populations, Proc. of Conference on Agricultural & Environmental Statistical Applications in Rome, Contr. Papers, pp XLVI-1,2, Rome, Italy.