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. \(D \subseteq R^k\), the need of a low budget sampling arises. Every unit \(u \in D\) figures as the \(u(x_1,x_2,...,x_k) \in D \subseteq R^k\), 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

\[
Z(u) = f(x_1,x_2,...,x_k) \in R, \quad u(x_1,x_2,...,x_k) \in D.
\]  

(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 \(f(x_1,x_2,...,x_k)\) subject -probably- to some parameters. It is well known that the mean value of \(Z(u)\) is given by

\[
\mu_z = \overline{Z} = \frac{1}{|D|} \iint_{D} f(x_1,x_2,...,x_k) \, dx_1 dx_2 ... dx_k,
\]

(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 \(\mu_z\), based only on a sample of points (units of \(P\)) taken from an arc \(C \subseteq D\) with dimension \(m\), \(1 \leq m < k\), 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 \(R^k\).

\(R^k\) We are going to draw elements only from the arc \(C=(C) \cup D\). If the form of \(f\) in (1.1) involves unknown parameters, then a resampling procedure may be used to provide point or (better) interval estimates 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 \(C=(C) \cup D\), 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 \(Z(u) = f(x_1,x_2,...,x_k)\) \(u \in D\). 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). 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 \(C=(C) \cup D\). The curve \((C)\) is a m-dimensional subspace of \(R^k, m<k\), i.e. \(C\) is the set of points (elements):

\[
(C) = \{x_1,x_2,...,x_k | x_m, x_{1}, x_2,...,x_{k-m} | x_j \in R, j=1,...,k-m \}
\]

and denote the measure (norm) of \(C\), as \(|C| \in R\).

Observe that arc \(C \subseteq R^k\). \(R^k\).

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

\[
\overline{Z}_c = \frac{1}{|C|} \iint_{C} f(x_1,x_2,...,x_k) \, dx_1 dx_2 ... dx_k
\]

(2.2)

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

\[
\Delta \overline{Z} = 0
\]

(2.3)

\[
\Delta \overline{Z} = \Delta \overline{Z} = \min \Delta \overline{Z}
\]

(2.3a)

\[
\frac{\Delta \overline{Z}}{\overline{Z}} \leq \frac{d}{d \in (0.00,a), a \leq 0.20}
\]

(2.3b)

STEP 6: Take a sample of points (e.g. systematic) on...
Nicolas: Sampling Designs Based on Subpopulations

the arc C and calculate the estimator \( \hat{Z} \) of \( Z = \mu_z \)

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
Ecology, 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
\( D = \{(x,y), 0 \leq x, y' \leq t \} \).

Answer: The proposed method gives, as a solution of
\( \Delta Z = 0 \), the line \( y = 0.75x \), \( 0 \leq x \leq t \). The
suitable arc (space) of this line is \( C = \{(x, 0.75x),
0 \leq x \leq t \} \), 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, 0.611t,
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
\( \hat{\mu}_1 = \bar{Z} = 0.2540002 \cdot a \cdot t^2 \).

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

Answer: The points are \( S_2 = \{x_i, x_i = \frac{r_i - t}{31},
0 \leq r_i \leq 31 \} \) and of course it is
\( Z = 0.75 \cdot a \cdot t^2 \). So we obtain
\( \hat{\mu}_1 = \bar{Z} = 0.2540322 \cdot a \cdot t^2 \).

Example 3.3: Take a systematic sample of size \( n \) in the
case described in example 3.1.

<table>
<thead>
<tr>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 sec</td>
</tr>
<tr>
<td>3 sec</td>
</tr>
<tr>
<td>5 sec</td>
</tr>
</tbody>
</table>

\( r \rightarrow 1 \mu \)

\( 9480900 \)
\( 9265400 \)
\( 8849000 \)

\( 6 \mu \)
\( 9141200 \)
\( 8933300 \)
\( 8532100 \)

\( 9 \mu \)
\( 8943100 \)
\( 8740000 \)
\( 8346500 \)

\( 12 \mu \)
\( 8749400 \)
\( 8550000 \)
\( 8116000 \)

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

\[ Z(t, r) = m_0 \cdot e^{-ar} = M_0 \cdot e^{-bt} \cdot e^{-ar} \]

\( m_0(r) = Z(t, r) = M_0 \cdot e^{-a(t+b)} \) \hspace{1cm} (3.3)

Answer: The points are \( S_n = \{x_i, x_i = \frac{r_i - t}{n - 1},
0 \leq r_i \leq n-1 \} \) and of course it is
\( Z = 0.75 \cdot a \cdot t^2 \). So we obtain
\( \hat{\mu}_1 = \bar{Z} = 0.25 \cdot a \cdot t^2 \).

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

Example 3.4: Take \( Z(u)=f(x,y) = (x^2 - y^2)^2 \), with
\( D = \{(x,y), x^2 + y^2 \leq a^2 \} \). Give an estimation for \( \mu = \frac{2}{3} \cdot a \)
(Farmakis, 1999.)

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

\[ (C) \quad y = 0.5287 \cdot a \quad \text{and} \quad (x, y) \in \mathbb{D} \]

A systematic sample of \( n = 101 \) points gives an
estimation of \( \mu \):

\[ \hat{\mu}_1(101) = 0.659368 \cdot a \]

Also for \( n = 201 \) points it is:

\[ \hat{\mu}_1(201) = 0.663101 \cdot a = \frac{2}{3} \cdot a \]

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

Example 3.5: The number of the microbes crowded, at
the moment \( t \), per every area \( ds=1\mu^2 \), is a function of
the distance \( r \) from the origin \( O(0,0) \). This
number is given by

\[ m_0(r) = M_0 \cdot e^{-ar}, \quad 0 \leq r \leq (x^2 + y^2)^{0.5} \leq 20, \]

0 ≤ t ≤ 10 \hspace{1cm} (3.1)

Also the number of the microbes at the origin at the
moment \( t = 0 \) is

\[ M_0 = m_0(0) = M_0 \cdot e^{-bt} \]

0 ≤ t ≤ 10 \hspace{1cm} (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.

\[ \ln Z(t, r) = \ln M_0 - a \cdot r - b \cdot t \] \hspace{1cm} (3.4)

After some calculations the results of regression
analysis are:

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

\[ \ln Z(t, r) = \ln M_0 - a \cdot r - b \cdot t \]

\[ M_0 = M_0 \cdot e^{-bt} \]

4 \hspace{1cm} (3.2)
Nicolas: Sampling Designs Based on Subpopulations

\[
\begin{bmatrix}
\ln \hat{M}_0 \\
\hat{a} \\
\hat{b}
\end{bmatrix} =
\begin{bmatrix}
16.12051 \\
0.00746 \\
0.02355
\end{bmatrix}
\]
i.e. \( m_1(r) \) in (3.3) is estimated as

\[ m_1(r) = Z(t, r) = 1.0024 \cdot 10^7 \cdot e^{-(0.00746 + 0.02355) r} \]  \hfill (3.5)

(b) The mean value of the number of microbes per \( \mu^2 \) in the points of \( D \) is now estimated by

\[ \hat{Z} = \hat{m} = \frac{1.0024 \cdot 10^7}{\| D \|} \int \int e^{-(0.00746 + 0.02355) r} dr \, dt 
= 8297180 \text{ microbes/} \mu^2 \]

with \( D = \{(r, t): 0 \leq r = (x^2 + y^2)^{1/2} \leq 20, 0 \leq t \leq 10 \} \)

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 \( (e): \lambda \cdot r, \lambda \in \mathbb{R} \), in order to calculate the value of \( \bar{Z}_C(r) = \bar{m}_{C, r}(r) = \frac{1.0024 \cdot 10^7}{20} \int_0^{20} e^{-(0.00746 + 0.02355) r} dr \) \hfill (3.6)

and to solve the equation (2.3), STEP 5, for the value of \( \lambda \), 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 \( \lambda = 0.513 \) approximately, i.e. we have the line

(\( e \)): \( t = 0.513 \cdot r \) \hfill (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 \( \bar{Z} \)

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

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

We will use the set \( C = \{(e) \cap D \} \) as the population from which we will draw off a sample of size \( n = 21 \) (systematic) to estimate the mean value of \( m_1(r) \).

Thus we have:

\[ \hat{m} = \hat{m}_{C, r}(r) = \frac{1.0024 \cdot 10^7}{21} \sum_{r=0}^{20} e^{-0.01954 r} \]

microbes/\( \mu^2 \)

\[ = 0.4773 \cdot 10^5 \]  \hfill (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 \( \bar{Z} = 8297180 \) microbes/\( \mu^2 \), the new estimation in (3.9) seems to be a good one, with a relative error (difference) 0.0657% to \( \hat{Z} = 8297180 \). 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


