Metodološki zvezki, Vol. 7, No. 1, 2010, 17-23

Some Notes on the Probability Space of Statistical Surveys

George Petrakos[1]

Abstract

This paper introduces a formal presentation of sampling process using principles and concepts from Probability Algebra and Information Theory. Under this model, any sampling scheme defines uniquely a probability measure, illustrated in various examples along with some applications in survey design and management.

1 Introduction – basic definitions

Let P be the target population of a statistical survey and R = $\{r_\nu, \nu = 1,2, \ldots N\}$ a relevant register in hand, consists of N individual statistical units. Regardless of the parameters of the selection process, all the possible outcomes concerning the elements of R comprise the set $\Omega_R = \{r_1^+, r_1^-, r_2^+, r_2^-, \ldots r_N^+, r_N^-\}$, where r_ν^+ denotes the presence of the ν^{th} unit while r_ν^- denotes its absence (Kullback, 1997). By $\mathscr{B} = \{E \subseteq \Omega_R\}$ we determine the set of all subsets E of $\Omega_{R,}$ called samples. Any selection process in Ω_R defines uniquely a probability measure p and furthermore any sample E can either have chances to appear (p(E) > 0) or not (p(E) = 0). Let us now consider a mapping on $\mathscr{B}$ ($\phi : \mathscr{B} \rightarrow \mathscr{E}$) such that,

$$\phi\,(\mathrm{E}) = \begin{cases} \mathrm{E}, & \mathrm{p(E)} > 0 \\ \varnothing, & \mathrm{p(E)} = 0 \end{cases} \tag{1.1}$$

[1] Dept. of Public Administration, Panteion University of Social and Political Sciences. Athens, GR **and** Agilis SA, Statistics and Informatics, GR Acadimias 96-100, 10677 Athens, Greece; george.petrakos@agilis-sa.gr

Thus we construct a non-empty set $\mathscr{E} \subseteq \mathscr{B}$ which, with the basic Boolean operations and a probability measure p which is strictly positive, normed and additive, form a probability algebra ($\mathscr{E}$, p) (Kappos, 1969). Therefore for any elements E in $\mathscr{E}$

(i) $p(E) > 0$ and $p(E) = 0$ iff $E= \varnothing$

(ii) $p(e) = 1$, where e is the unit in $\mathscr{E}$

(iii) $p(E_1 \cup E_2) = p(E_1) + p(E_2)$ if $E_1 \cap E_2 = \varnothing$

Any element in $\mathscr{E}$ different than $\varnothing$ and e is called possible sample. We also consider N+1 classes $\mathsf{S}(n) \subseteq \mathscr{B}$, n =0,1,2,…N such that

$\mathsf{S}^{(n)} = \{ (r_1^k, r_2^k, \dots r_N^k), \quad \sum_{\nu=1}^{N} I(r_\nu^k) = n \}$, where k = {+, -} and

$$I(r_i^k) = \begin{cases} 0, & k = - \\ 1, & k = + \end{cases}$$

which contains all subsets of $\mathscr{B}$, where n appearances of statistical units occur.

$\forall$ n, $1 \leq n \leq N$, the class $\mathsf{S}^{(n)}$ contains $\binom{N}{n}$ subsets $\mathsf{S}_i^{(n)}$, $i \in I_n = \{1, 2, \dots \binom{N}{n}\}$

By applying ϕ on $\mathsf{S}(n) \subseteq \mathscr{B}$, we construct a non empty set $\mathsf{S}n$

$$\phi: \ \mathsf{S}^{(n)} \rightarrow \mathsf{S}^{n} \ \therefore \phi(\mathsf{S}_i^{(n)}) = \begin{cases} S_i^{(n)}, \ p(S_i^{(n)}) > 0 \\ \varnothing, \ p(S_i^{(n)}) = 0 \end{cases}, \quad i \in I_n \tag{1.2}$$

Under the probability algebra ($\mathscr{E}$, p) defined by a chosen sampling process, the class $\mathsf{S}n$ has the following properties (inherited by $\mathscr{E}$)

(i) $p(\mathsf{S}_i^n) > 0$, $i \in I_n(s)$

(ii) $p(\cup \mathsf{S}_i^n) = 1$, $i \in I_n(s)$

(iii) $p(\mathsf{S}_i^n \cup \mathsf{S}_j^n) = p(\mathsf{S}_i^n) + p(\mathsf{S}_j^n)$, $\forall$ (i, j) $\in I_n(s) \times I_n(s)$ with $i \neq j$

where, $I_n(s) \subseteq I_n$ the subset of indices for which $p(\mathsf{S}_i^{(n)}) > 0$ and $e = \cup \mathsf{S}_i^n$ $i \in I_n(s)$, the unit, with p(e) = 1.

This basic set of notions and definitions introduces a more algebraic approach to measurable sample designs than the analytical ones (Särndal et all, 2003) which are focusing on the estimation of various parameters. This algebraic approach seems to handle multiple sampling procedures, like multiple recapture designs, more efficiently.

2 Application to various sampling schemes

In a single sample process it can be shown that $S_i^n \cap S_j^n = \varnothing$, $\forall$ (i, j) $\in$ $I_n(s)$ x $I_n(s)$ with i≠j. The probability that two different samples will be drawn in a single sampling process is zero, therefore p($S_i^{(n)} \cap S_j^{(n)}$) = 0 and the only event in ($\mathscr{E}$, p) with probability 0 is the empty set, $\varnothing$. There are sampling schemes where $I_n(s) \subset I_n$ (strictly), i.e. in stratified random sampling where only the $S_i^{(n)}$s that satisfy the proportional to strata restriction meet with property (i), while for the rest it holds that p($S_i^{(n)}$) = 0, i$\in I_n - I_n(s)$. On the other hand, in a simple random sampling $I_n \equiv I_n(s)$, since all $S_i^{(n)}$, i$\in$ I_n satisfy property (i). The above concepts can also be applied to multiple sampling procedures. In this type of sampling, both r_v^+ and r_v^- are present in the sample, in different stages of course. We will examine the form of the event space Ω_R and the class S^n, for sampling with replacement and multiple recapture sampling.

Sampling with replacement. Sampling from N statistical units by choosing one unit each of the n(sample size) times and put it back in the population before the next trial is a process that corresponds to an event space Ω_R such that:

Ω_R = {$r_v^k(n)$} with ν = 1,2,…,N k={+,-} and n = 1,2,… where $\sum_{\nu=1}^{N} I[r_\nu^k(n)] = 1$, ∀ n and a probability algebra ($\mathscr{E}$, p) is defined based on $S^n = S^1$ x S^1x…x $S^1 = \underset{n}{X} S^1$, where S^1 is the basis for an SRS of size 1.

Multiple recapture. In a multiple recapture experiment run in a population of size N (usually unknown), the sample space is expanded over the discrete time of trials (t=1,2,…T). If the population is closed for this time period, the sample space is:

$\Omega_R(T)$ = {$r_v^k(t)$} with ν = 1,2,…,N, k={+,-} and t = 1,2,…,T. When the population size changes in the different points of time (open population), the sample space is:

$\Omega_R(T)$ = {$r_v^k(t)$} with ν = 1,2,…,N(t), k={+,-} and t = 1,2,…,T. The basic class is $S^T = S^{X_1} x S^{X_2} x ... x S^{X_T} = \underset{t}{X} S^{X_t}$, where $X_t \in$ {0,1,…,N(t)} a discrete random variable with elements $S_i^T = S_{i_1}{}^{X_1} x S_{i_2}{}^{X_2} x ... x S_{i_T}{}^{X_T}$ with I_t =1, 2, … $\binom{N}{X_t}$, t = 1,2,…,T.

3 The probability space

Under a pr. algebra ($\mathscr{E}$, p) a class S^n is uniquely defined and contains all the possible samples and only them. This class forms a basis for the construction of all events in $\mathscr{E}$. Any event E$\in$ $\mathscr{E}$ can be constructed by using one or more basic samples S_i^n and expressed as a union of these S_i^n, based on the fact that any possible event related to the sampling process can be realized by unions of samples S_i^n, $I_n(s)$.

It can be easily shown that $\mathscr{E}$ is closed under the basic set operation. For that, let us consider E_1, $E_2 \in$ $\mathscr{E}$ as unions of some S_i^n, such that:

$E_1 \in \mathscr{E} \Rightarrow E_1 = \bigcup_i S_i^{(n)}$, $E_2 \in \mathscr{E} \Rightarrow E_2 = \bigcup_j S_j^{(n)}$, for some i, j $\in$ $I_n(s)$,. Then

$E_1 \cup E_2 = \bigcup_i S_i^{(n)} \cup \bigcup_j S_j^{(n)} = \bigcup_d S_d^{(n)} \in \mathscr{E}$ where d is such that $S_d^{(n)}$ belongs either to $\bigcup_i S_i^{(n)}$ or $\bigcup_j S_j^{(n)}$ and $E_1 \cap E_2 = \bigcup_g S_g^{(n)} \in \mathscr{E}$ where g is such that $S_g^{(n)}$ belongs both to $\bigcup_i S_i^{(n)}$ and $\bigcup_j S_j^{(n)}$. If there is no g such that $S_g^{(n)}$ belongs to both of the unions above, then $E_1 \cap E_2 = \varnothing$ and the two events are mutually exclusive. These properties can be easily extended for any finite set of events E_i. Moreover, the above defined possible event E_2 contains another possible event, noted as $E_1 \subseteq E_2$ when

$$\bigcup_i S_i^{(n)} \subseteq \bigcup_d S_d^{(n)} \text{ , } i\in I,\ d\in D, \text{ or equivalently } I\subseteq D.$$

Let us now illustrate the above with a couple of examples:

Example 1 Let N = 4 and n = 3. Then S^3 is a class of four basic sets, namely , $S_1^3 = \{r_1^-, r_2^+, r_3^+, r_4^+\}$, $S_2^3 = \{ r_1^+, r_2^-, r_3^+, r_4^+\}$, $S_3^3 = \{ r_1^+, r_2^+, r_3^-, r_4^+\}$, $S_4^3 = \{ r_1^+, r_2^+, r_3^+, r_4\}$. The event of the presence of the first two individuals which can be noted by $E_{12} = \{r_1^+, r_2^+\}$ can be expressed as a union of basic sets, $E_{12} = S_3^{(3)} \cup S_4^{(3)}$. In other words, the event E_{12} occurs when at least one of the basic events in which the first two individuals are present occurs.

Remark Someone can argue that in the example above that $B_{12} = S_3^{(3)} \cap S_4^{(3)}$, which in terms of point set theory seems correct, since $\{ r_1^+, r_2^+, r_3^-, r_4^+\} \cap \{ r_1^+, r_2^+, r_3^+, r_4^-\} = \{r_1^+, r_2^+\}$. However, in our treatment under the given sampling scheme, $\{r_1^+, r_2^+\} \equiv \{r_1^+, r_2^+, r_3^k, r_4^k\}$, k=+,- which explains why $B_{12} = S_3^{(3)} \cup S_4^{(3)}$

Example 2 Let N = 3 so R = $\{r_1, r_2, r_3\}$. If they are placed in an orthogonal space in 3-D taking values of 0 and 1 for non-appearance and appearance respectively, we have the following transformation:

$[S^{(0)}]$: $(0,0,0) \rightarrow (r_1^-, r_2^-, r_3^-)$
$[S^{(1)}]$: $(1,0,0) \rightarrow (r_1^+, r_2^-, r_3^-)$, $(0,1,0) \rightarrow (r_1^-, r_2^+, r_3^-)$, $(0,0,1) \rightarrow (r_1^-, r_2^-, r_3^+)$
$[S^{(2)}]$: $(1,1,0) \rightarrow (r_1^+, r_2^+, r_3^-)$, $(1,0,1) \rightarrow (r_1^+, r_2^-, r_3^+)$, $(0,1,1) \rightarrow (r_1^-, r_2^+, r_3^+)$
$[S^{(3)}]$: $(1,1,1) \rightarrow (r_1^+, r_2^+, r_3^+)$

which produce all possible samples. For n=2, we have 3 orthogonal vectors $S_1^{(2)}$, $S_2^{(2)}$, $S_3^{(2)}$ which are a basis for some sampling schemes where 2 out of 3 are selected.

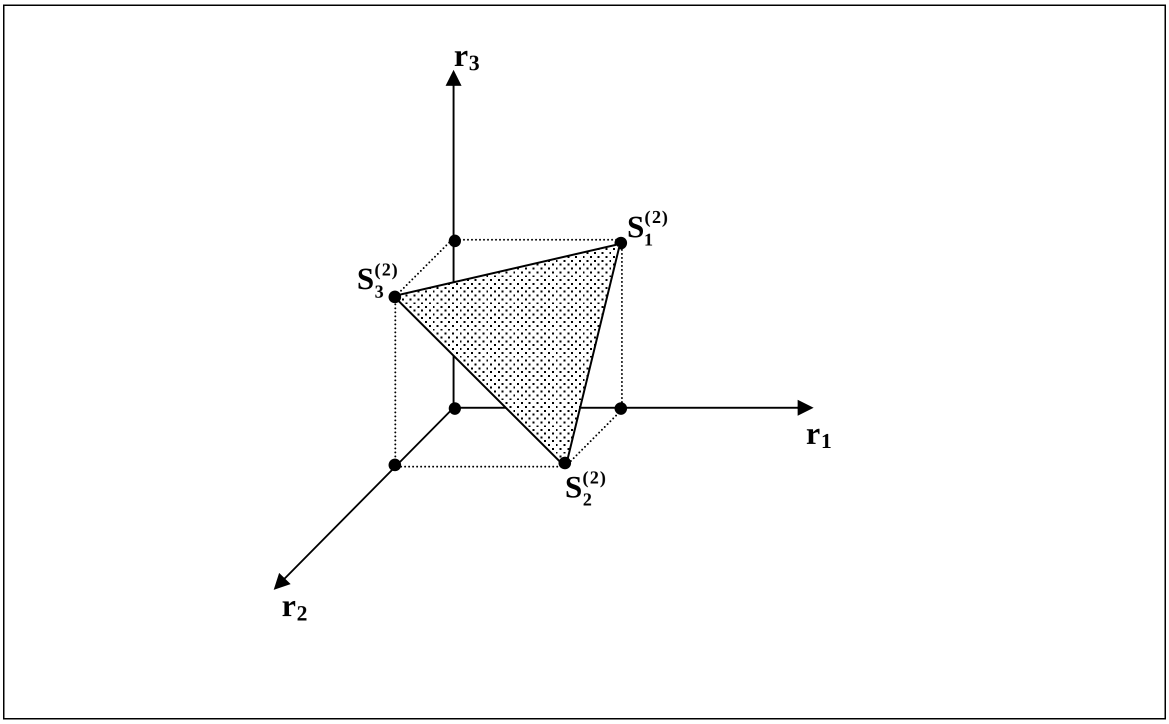

Figure 1: 3-D orthogonal space.

Any measure f_i applied to r_i can associate a measurable function $\varphi^{(n)} =(f_i)$ i=1,2,…n to a class S^n and therefore a value $\varphi^{(n)}$ (S_i^n) to each basic sample. Considering the probability measure p in S^n mathematical expectation

$$E(\varphi^{(n)}) = \sum_{i} p(S_i^n)\, \varphi^{(n)} (S_i^n) \qquad (3.1)$$

where $\varphi^{(n)} = \{ f_1(r_1), f_2(r_2),\ldots f_n(r_n)\}$ and $p(S_i^n)$, $\sum_{i} p(S_i^n) =1$, is located at the barycenter of the polytope formed by S^n (Petrakos, 2000) and expresses a mean value of $\varphi^{(n)}$ before the sample is drawn. An interesting application of this approach is the determination and application of a cost function $C^{(n)} = \{c_1(r_1), c_2(r_2),\ldots c_n(r_n)\}$, where c_i's variation is due to corresponding r_i's costly characteristics (access, distant location, etc). Then the cost of a sample S_i^n is $C_i = C^{(n)} I(S_i^n)'$, where $I(S_i^n)$ is a n-dim vector with ones for the corresponding r_i^+ ‘s and zeros for the r_i^- ‘s. Finally the expected cost of the sampling process estimated in the design phase will be

$$E(C^{(n)}) = \sum_{i} p(S_i^n)\, C^{(n)}\, I(S_i^n)' \qquad (3.2)$$

4 Conclusions

A probability algebra model has been introduced in order to describe the data collection process in a statistical survey. Its sufficiency, efficiency and simplicity was tested and proved over different sampling schemes. Future research can adapt this model to more complicated and realistic sampling schemes, incorporating cost and non-response to the design of a statistical survey. From a theoretical point of view, this model can be viewed and further studied as an application of group theory. In both cases, this paper aspires to provide some basic ideas for substantial research.

Acknowledgements

The author is grateful to the associate editor of the journal and to the referees for their constructive comments. The author would also like to thank Mr. George Maniatis for reviewing the final version of this paper.

References

[1] Cochran, W. (1977): *Sampling Techniques*. New York: J. Wiley & Sons.

[2] Kappos, D. (1969): *Probability Algebras and Stochastic Spaces*. Monograph in Probability and Mathematical Statistics. London: Academic Press.

[3] Kullback, S. (1997): *Information Theory and Statistics*. New York: Dover Publ. Inc.

[4] Petrakos, G. (2000): The topological foundation and some properties of the mixed estimator. *Computational Statistics*, **15**, 109-114

[5] Särndal, C., Swensson, B., and Wretman, J. (2003): *Model Assisted Survey Sampling*. New York: Springer.