STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 313—328

MODELING OF THE INTERNET SAMPLE

Getka-Wilczyńska Elżbieta[1]

ABSTRACT

In the paper we present the following ideas for the Internet mediated research – respondents' arrival process as a stream of events and a pure birth process and cut models of it for exponential distribution to interpret and analyse some stochastic properties of the Internet data collection process.

Key words: Internet sample, population, Markov process, pure birth process, model.

1. Introduction

Statistical research concerning sampling selection can be divided into representative surveys based on the probability sample (the sample is random and all statistical units should have a strictly positive probability of being selected to the sample) and surveys based on the non-probability sample. After choosing the kind of the sample selection the next stage of the survey is data gathering. In all surveys the data is collected by using an immediate interview, a telephone interview, a mobile phone, a post or by using a computer and in recent years an interview over the Internet (see www.WebSM.org.; www.aapor.org).

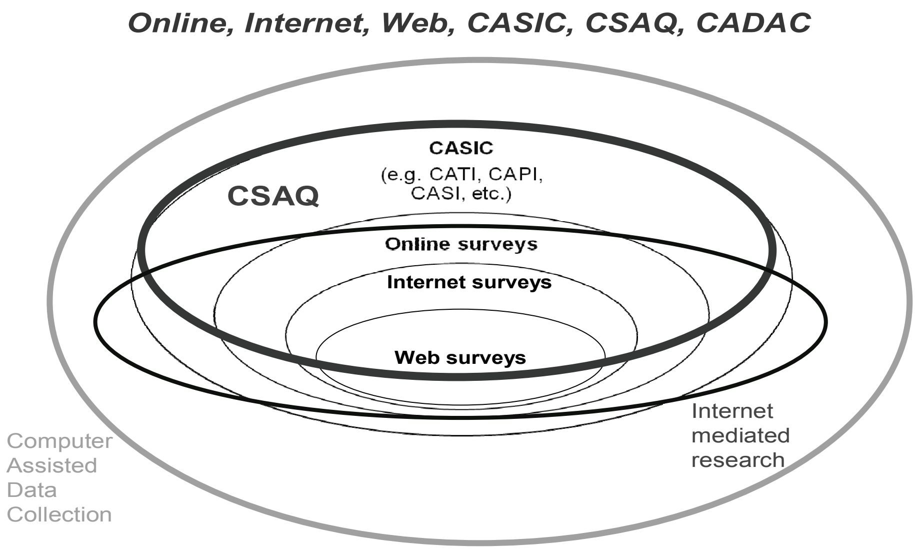
The Internet mediated research is in a process of intensive development and the key characteristic of it is its diversity. Collecting data in these surveys is useful for major corporations, small business, micro-business, academic researches and students, non-profit organizations, political organizations, individuals, government, private research societies and statistical agencies. The Internet mediated surveys have several advantages, such as low cost of collecting information, integration across devices, modes and processes, speed of the data transmission and opportunity to monitor it. Moreover, the computerized nature of Web surveys facilitates conducting experiments. The usage of the electronic questionnaire in the Internet mediated survey makes the interview more efficient, lowers the workload of the respondents and controls the response quality: an elimination of errors during data transcription, an implementation of advanced

[1] Warsaw School of Economics, Institute of Econometrics, Division of Mathematical Statistics, Al. Niepodległości 162, 02-554, Warsaw, Poland, Elzbieta.Getka-Wilczynska@sgh.waw.pl

features: automatic skips and branching, a randomization of questions. The first graphic browser (NCSA Mosaic) was released in 1992, with Netscape Navigator following in 1994 and Internet Explorer in 1995. The first published papers on Web surveys appeared in 1996. Since then, there has been a virtual increasing tendency of interest in the Internet generally, and World Wide Web specifically, as a tool of data collection. A special portal WebSM (www.WebSM.org) – Web survey methodology web site is a website dedicated to the methodology of Web surveys. It has been supported by the EU since 2002 and it includes bibliography lists and software database.

Figure 1 shows competing, overlapping, complementing, synonymous terms: Web surveys, Internet surveys, Internet mediated surveys, Internet mediated research, on-line surveys, CSAQ (computerized self-administrated questioners), CASIC (computer assisted survey information collection), Telesurveys, CADAC (computer assisted data collection).

Figure 1. Modes of data collection by using computer: Online, Internet, Web, CASIC, CSAQ, CADAC (Vehovar, 2007)



But the basic problem in the surveys over the Internet is concerned with collecting and analysing data sets according to classical methods of the sampling theory and statistical inference based on the probability sample.

Generally, differences between representative surveys and Internet mediated researches rely on the following aspects. In representative surveys based on the probability sample the frame of sampling or registers is available, respondents are drawn to the sample by a statistician according to sampling design (sampling scheme) and the methods of sampling theory and statistical inference are applied to data analysis (e.g. Särndal et al. (1992), Tillě (2006)). If in representative surveys respondents are randomly selected to the probability sample and instead of the traditional modes of interview an electronic questionnaire is used, then it is only one of modes of data collection (one except – complete survey) and the correct usage of this data collection tool requires a suitable survey methodology.

In the most of the Internet mediated researches, the basic problem is concerned with imperfect frames, access to the computer (or the Internet), the self-selection error, issues on the representation, especially with regard to more general population and an application of statistical inference methods to the survey based on the non- probability sample.

If well-defined frames (that is, the current and complete list of identifiable respondents with access to the computer, the Internet) are not available in form suitable for sampling or do not exist, then drawing the probability samples is not possible. For example, drawing the sample is not possible because the preparation of the frame may not be possible – e.g. for users of computers, users of the Internet and different subpopulations of Internet users – users of Skype, Internet portals or Internet websites, consumers of company products, different virtual social groups as focus and chat groups, blogs and bulletin boards and so on (usually the size of the population is not known, the elements of the population are not identified). Sometimes drawing the probability sample is possible (e.g. there exists the frame of households), but there does not exist the frame of users with the access to the computer; there exists the data base of the population (e.g. the administrative data base of scientists), but large costs of preparing of the frame exclude the survey of this population based on the probability sample; the surveyed population is given by the frame, the register or the list, all units have the access to the electronic questionnaire (e.g. small business), all units of the frame are identified and then the frame – based Internet survey (the complete survey of the population) is possible.

The mentioned cases do not describe all possibilities, but generally, if the frame does not exist or is not available then drawing probability sample is impossible and units of the surveyed population cannot be included in the sample according with the classical sampling methods which are useful when it is possible to examine all the units of a finite population.

In these surveys the respondents who belong to the population of interest (the size of the population is known or not) are not randomly selected to the sample, but they participate in the survey according to their subjective decision and data are collected in an uncontrolled way. This is a source of the self-selection error and problem with representativeness of these surveys. The methods of the sampling theory based on the probability sample cannot be used for the data from such samples because the inclusion probabilities are not known (the exception – the complete survey) and statistics are calculated on the basis of the Internet data referring only to the population surveyed.

In theory and practice of the Internet survey two approaches to deal with this problem are identified by Couper and Miller (2008). The first one, the design approach, is mainly associated with the use of panels and is based on the idea of building probability – based on Internet panels by using other methods for sampling and recruitment and, if necessary, providing Internet access to those without. This approach is applied, e.g. by Knowledge Networks in the USA and CentERdata's MESS panel in the Netherlands. The second one, the model based approach, begins with a volunteer – in panel of Internet users, and attempts to correct representation biases using e.g. propensity score adjustment (Lee (2006)) or some other weighting method for assessing Web panel quality (Rosenbaum and Rubin (1983), Callegaro and Disogra (2008), Lee and Valliant (2009), Schonlau et al. (2009)). In both approaches the methodology of sampling theory and statistical inference to data analysis is used. The other interesting proposition is an application of the dynamic theory of decision making and the decision field theory to theoretical explanation of survey behaviour (Galesic (2006)).

Generally, we propose using stochastic processes (the birth and death processes) to study (the description and interpretation, modelling and possible analysis) models of the population and data collecting in the Internet mediated research. In this approach a stochastic nature of events which appear during the Internet mediated survey is presented in following way.

Firstly, when the process of the Internet data collection is treated as a process of registering questionnaires on the server, we assume that the moments of recording the questionnaires form a random sequence of arrivals (or responses) of these respondents who took part in the survey at these moments. This sequence can be modelled in different ways. Moreover, the respondents who took part in the survey form a random subset of the surveyed population called an uncontrolled sample, the Internet sample or the self-selection sample, and the size of this sample is defined as a counting process built on the moments of recording the questionnaires.

Secondly, the Internet mediated survey with the process of the Internet data collection is considered as a life test of the population of the finite size treated as coherent system and the basic characteristics of the population lifetime for proposed models are defined and calculated by using Markov methods (Getka-Wilczyńska (2009)).

The paper is structured as follows. Section 2 contains a notion and a definition of the size of the uncontrolled sample as a counting process. In section 3 the pure birth process and its properties as well as particular cases: the Poisson process and two models for the population of finite size are presented.

2. Preliminaries

We assume that the Internet mediated survey begins at the moment $t = 0$, when the electronic questionnaire is put on the website and the survey is conducted for the time $T > 0$. A set $\{u_1, u_2,...,\}$ denotes the population of potential respondents and the respondents fill in questionnaires independently. By $\tau_0 \le \tau_1 \le \tau_2 \le ...$ are denoted the successive random moments of recording the questionnaires on the server after an initial moment $t = 0$, $\tau_0 = 0$. In this case, τ_k, $k \ge 1$, is interpreted as the random moment of an arrival (or a response) of the respondent u_k, $k \ge 1$, who belongs to the population of potential respondents.

For each $n \ge 1$ the population of finite size *n* is surveyed and the size of the population is random. In this case τ_k, $k = 1,2,...,n, n \ge 1$, is interpreted as the random moment of an arrival (or a response) of this respondent u_j, $j = 1,2,...,n, n \ge 1$, belonging to the population of size n, $n \ge 1$, who took part in the survey as the k th.

We assume that X_k, $k = 1,2,...,n$, $n \ge 1$, are nonnegative random variables with distribution function $F_k(t) = P(X_k \le t)$, for $t \ge 0$, $k = 1,2,...,n$, $n \ge 1$, the probability density function $f_k(t) = F_k'(t)$, $F_k(t) = \int_0^t f_k(x)dx$.

Random variable X_k, $k = 1,2,...,n$, $n \ge 1$ is interpreted as a waiting time of the arrival (the response) of the *k*th respondent up to time *t* when the size of the uncontrolled sample is determined or a lifetime of *k*th element of the population of interest up to time *t* when the population lifetime is considered.

2.1. Random size of the uncontrolled sample

If the process of Internet data collection is considered as a process of registering questionnaires on the server in a fixed interval of the time $T > 0$ (the time of the survey conducted) then the size of the uncontrolled sample at the moment $t \ge 0$ equals total number of respondents' arrivals up to moment $t \ge 0$ and is defined as a counting process $\{N(t), t \ge 0\}$ built on the sequence $(\tau_0, \tau_1, \tau_2....)$ as follows.

We assume that no arrival has occurred up to time $t = 0$.

The number of respondents' arrivals up to time $t \ge 0$ is given by

$$N(t) = card\{k \ge 1 : \tau_k \le t\} = \max\{k \ge 1 : \tau_k \le t\}$$

and satisfies the conditions of definitions 1.1-3.1 below (Resnick (1998)).

Definition 2.1 The process $N = \{N(t) : t \ge 0\}$ is called a counting process (stream of events) if for all $t, h \ge 0$: 1. $N(0) = 0$ 2. $N(t) \in \mathrm{N}$ 3. $N(t) \le N(t+h)$

The increments $N(t+h)-N(t)$ of the process N models the number of arrivals occurring in the interval $(t,t+h]$. Realizations of counting process are monotonically non-decreasing and right-continuous functions. The process N is completely defined, if for $n \geq 1$ and any nonnegative numbers $t_1, t_2,..., t_n$ a probability distribution of a random vector $(N(t_1), N(t_2),..., N(t_n))$ is determined.

Definition 2.2 (equivalent to definition 2.1). Let $\tau_1, \tau_2,....,$ be successive moments of occurrence of events, $\tau_{k-1} \leq \tau_k$ for $k \geq 1$ and $\tau_0 = 0$.

Random variables $T_k = \tau_k - \tau_{k-1}$, $k \geq 1$ denote the time between the kth and (k-1)st records of questionnaires and $\tau_k = \sum_{j=1}^{k} T_j$, $k \geq 1$, $T_0 \equiv 0$.

The process $N = \{N(t): t \geq 0\}$ is called a counting process (stream of events), if for all $n \geq 1$ and any nonnegative integer numbers $t_1, t_2,..., t_n$ a probability distribution of a random vector $(T_1, T_2,..., T_n)$ is determined.

Definition 2.3 Let $(T_k)_{k\geq 1}$ be a sequence of nonnegative random variables.

We denote $\tau_n = T_1 + T_2 + ... + T_n$ for $n \geq 1$, $\tau_0 = 0$

$$N(t) = card\{k \geq 1 : \tau_k \leq t\} = \max\{k \geq 1 : \tau_k \leq t\}.$$

A sequence $(\tau_n)_{n\geq 0}$ is called a renewal stream (renewal sequence).

The process $N = \{N(t): t \geq 0\}$ is called a renewal process and is completely defined if for each $n \geq 1$ and any nonnegative numbers $t_1, t_2,..., t_n$ a probability distribution of a random vector $(T_1, T_2,..., T_n)$ is determined.

Definitions 2.1–2.3 determine an infinite stream of respondents' arrivals i.e. for any integer number $M \geq 0$ exists the number $t \geq 0$, so that $\Pr(N(t) \geq M) > 0$.

For each $t \geq 0$ the value of the counting variable $N(t)$ is called the size of the uncontrolled sample (the Internet sample) at the moment t. Because the relation between the sequence $(\tau_0, \tau_1, \tau_2....)$ and the process $\{N(t): t \geq 0\}$ is given by $\{N(t) = n\} = \{\tau_n \leq t < \tau_{n+1}\}$ hence

$$P_n(t) = \Pr\{N(t) = n\} = \Pr\{\tau_n \leq t < t_{n+1}\} = \left\{ \sum_{k=1}^{n} T_k \leq t < \sum_{k=1}^{n+1} T_k \right\}$$

for $n \geq 0, t \geq 0$.

A general computational expression for the probabilities $P_n(t)$ is impossible because the interdependences $(T_n)_{n\in N}$ between the successive arrivals (responses) are not specified.

If we assume that $(T_n)_{n\in N}$ is the sequence of independent random variables we have the following possibilities:

1. $F_k(t) = P(T_k \le t)$, $k \ge 1$
2. $F_k(t) = F(t)$ for $k \ge 2$, $F_1(t) = P(T_1 \le t)$, the general renewal process
3. $F_k(t) = F(t)$, $k \ge 1$, the renewal process
4. $F_k(t) = F(t) = 1 - e^{-\lambda t}$, $\lambda > 0, t \ge 0$, $k \ge 1$, the Poisson process, the simplest renewal process.

Below are presented two cases of the infinite stream of the respondents' arrivals: a homogenous pure birth process (1) and a particular case of it – a homogenous Poisson process (4) as well as some cut models of it for the population of finite size.

3. The pure birth process

During Internet mediated surveys we observe arrivals of the respondents belonging to the population $\{u_1, u_2, ...,\}$ at the moments τ_k, $k \ge 0$ and for each $t \ge 0$ the process $N = \{N(t): t \ge 0\}$. The random variable $N(t)$ (equals the number of arrivals up to time t), takes value 0,1,2,…

The Internet sample during the survey forms a random set of these respondents who participate in the survey. We say that the Internet sample is in the state E_n when $N(t) = n$ and $E = \{E_0, E_1, E_2, ...\}$ is a finite or a countable set of possible states of the Internet sample (E_0 denotes the state of the Internet sample in which $N(t) = 0$, E_1 denotes the state of the Internet sample, in which $N(t) = 1$,…. and so on). The transition from state E_n to state E_{n+1} occurs at the moment τ_n, $n \ge 1$, and denotes an increase of the size of the Internet sample by one. The Internet sample passes successively through the states E_0, E_1, E_2...... In the state $E_n, n \ge 0$ it is for the time $T_{n+1}, n \ge 0$, respectively, with the distribution $F_{n+1}(t) = P(T_{n+1} \le t)$ and next passes to the state E_{n+1} with the probability one.

According to the description stated above we postulate that the probability of the arrival of the respondent occurring at a given instant of time depends upon the number of the arrivals which have already occurred. It means that if during the interval $(0, t)$ n arrivals occur, then the probability of the new arrival in the interval $(t, t+h)$ is equal to $\lambda_n h + o(h)$, where $\lim\limits_{h \downarrow 0} \frac{o(h)}{h} = 0$ and the process $N = \{N(t) : t \geq 0\}$ is characterized by a sequence of positive numbers λ_0, λ_1, ...

Formally, the process $N = \{N(t) : t \geq 0\}$ is a Markov point process built on the sequence τ_n, $n \geq 0$, a homogeneous pure birth process. In this process the jumps intensity at any time *t* depends on the number of jumps before time *t* and does not depend on *t*. The process satisfies the following postulates:

1. N has the Markov property
2. $\Pr(N(h) = 1) = \Pr(N(t+h) - N(t) = 1 | N(t) = k) = \lambda_k h + o_{1,k}(h), \quad k \geq 0$, $h \to 0+$,
3. $\Pr(N(h) = 0) = \Pr(N(t+h) - N(t) = 0 | N(t) = k) = 1 - \lambda_k h + o_{2,k}(h), \quad k \geq 0$,
4. $N(0) = 0$,
5. $\Pr(N(t+h) - N(t) < 0 | N(t) = k) = 0, k \geq 0$.

The left sides of 2 and 3 are $P_{k,k+1}(h)$ and $P_{k,k}(h)$ respectively, so that $o_{1,k}(h), o_{2,k(h)}$ do not depend upon *t*. Because the size of the Internet sample is defined by

$$N(t) = card\ \{n \geq 1:\ \tau_n \in [0, t]\} = \max\left\{ n \geq 1 : \sum_{j=1}^{n} T_j \leq t \right\}$$

we are interested in the derivation of $P_n(t) = \Pr(N(t) = n), n \geq 0$.

The system of Kolmogorov differential equations satisfied $P_n(t) = \Pr(N(t) = n)$ for $t \geq 0$ is given by $P_0'(t) = -\lambda_0 P_0(t)$ $P_n'(t) = -\lambda_n P_n(t) + \lambda_{n-1} P_{n-1}(t)$, $n \geq 1$, with boundary conditions $P_0(0) = 1, \quad P_n(0) = 0, n > 0$. (e.g. Feller, t. I. p. 370, 1977; the solution, t. II, p. 431, 1978; Gnedenko et al. (1968)).

The first equation can be solved immediately and yields $P_0(t) = \exp(-\lambda_0 t) > 0$.

Since T_k is the time between the *(k-1)*th and *k-th* records (arrivals, responses), so that $P_n(t) = \left\{ \sum_{k=1}^{n} T_k \leq t < \sum_{k=1}^{n+1} T_k \right\}$ and $\tau_n = \sum_{k=1}^{n} T_j$ is equal to the time at which the *n*th record of the questionnaire (the arrival, the response) occurs. Therefore, $\Pr(T_1 \leq t) = 1 - \exp(-\lambda_0 t)$. The postulates 2-5 imply that random variable $T_k, k \geq 2$ has exponential distribution with parameter λ_{k-1} and T_k's are mutually independent.

Therefore, the characteristic function of $\tau_k = \sum_{j=1}^{k} T_j$ is given by

$$\varphi_{\tau_k}(t) = E(\exp(-it\,\tau_k)) = \prod_{j=1}^{k} E(\exp(-it\,T_j)) = \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\lambda_{j-1} + it}$$

$$\text{and} \quad Ee^{-\alpha\tau_k} = \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\lambda_{j-1} + \alpha} \quad \text{for } \alpha \neq 0 .$$

For a specific set of $\lambda_k \geq 0$ the solutions of the system of differential equations is given by the following formulae

$$P_k(t) = \lambda_{k-1} \exp(-\lambda_k t) \int_0^t \exp(\lambda_k x) P_{k-1}(x) dx \, , \; k = 1,2,\dots .$$

Hence, all $P_k(t) \geq 0$ for any $k \geq 0$ and $t \geq 0$.

But there is a possibility that for some sequences $\lambda_0, \lambda_1, \dots \sum_{k=1}^{\infty} P_k(t) < 1$. The intuitive argument for this fact is as follows: if the sum $\sum_{k=1}^{\infty} P_k(t)$ interpreted as the probability of the finite number of changes of the population states (records, arrivals, a response) up to the time t, then the difference $1 - \sum_{k=1}^{\infty} P_k(t)$ may be treated as the probability of the infinite number of the changes of the population states up to time *t* and, the process $N = \{N(t) : t \geq 0\}$ may explode, i.e. with positive probability $N(t) = \infty$. The necessary and sufficient conditions excluding the possibility of explosion is the equality (Feller, t. I, p. 373, 1977; t. II, p.431, 1978) $\sum_{k=0}^{\infty} \frac{1}{\lambda_k} = \infty$.

More formal argument for this result is as follows: the time T_k between consecutive arrivals (responses) has exponential distribution with parameter λ_k. Therefore, the quantity $\sum_{k=0}^{\infty} \frac{1}{\lambda_k}$ equals the expected time before the populations become infinite. By comparison, $1 - \sum_{k=1}^{\infty} P_k(t)$ is the probability that $N(t) = \infty$.

If $\sum_{k=0}^{\infty} \frac{1}{\lambda_k} < \infty$ the expected time for the population to became infinite is finite. It is then plausible that for all $t > 0$ the probability that $N(t) = \infty$ is positive.

If all $\lambda_0, \lambda_1, \dots \lambda_{n-1}$ are distinct and positive numbers then the finite dimensional distribution of the process $\{N(t) : t \geq 0\}$ is given by formulae

$$P_k(t) = \Pr(N(t) = k) = \left(\prod_{j=1}^{k} \lambda_{j-1}\right) \sum_{i=1}^{k} \frac{1 - e^{-t\lambda_{i-1}}}{\lambda_{i-1} \prod_{j=1, j\neq i}^{k} (\lambda_{j-1} - \lambda_{i-1})} -$$

$$- \left(\prod_{j=1}^{k+1} \lambda_{j-1}\right) \sum_{i=1}^{k+1} \frac{1 - e^{-t\lambda_{i-1}}}{\lambda_{i-1} \prod_{j=1, j\neq i}^{k+1} (\lambda_{j-1} - \lambda_{i-1})}$$

for $k = 1,2,\dots,n-1$ and $P_0(t) = \Pr(N(t) = 0) = \Pr(T_1 > t) = e^{-\lambda_1 t}$ for $k = 0$.

If $\lambda_n = 0$ we have $P_n(t) = \Pr(N(t) = n) = \left(\prod_{j=1}^{n} \lambda_{j-1}\right) \sum_{i=1}^{n} \frac{1 - e^{-t\lambda_{i-1}}}{\lambda_{i-1} \prod_{j=1, j\neq i}^{k} (\lambda_{j-1} - \lambda_{i-1})}$.

Finally, for this model the size of the uncontrolled sample is given by definition.

Definition 3.1 For $n \geq 1$ the size of uncontrolled sample until the moment $t \geq 0$ is given by

$$N(t) = card\ \{n \geq 1:\ \tau_n \in [0,t]\} = \max\left\{ n \geq 0 : \sum_{k=1}^{n} T_n \leq t \right\},$$

where $\tau_{k-1} \leq \tau_k$ for $k \geq 1$ are the successive moments of questionnaires record, $\tau_0 = 0$, $T_k = \tau_k - \tau_{k-1}$ for $k \geq 1$ are independent random variables with exponential distribution $Exp(\lambda_k)$.

The homogeneous pure birth process model is very interesting because it gives great flexibility for modelling the distribution of the number of arrivals as well as its evolution in time by proper choice of the sequence (λ_k). Different models determined by choosing specific sequence (λ_k) are given below.

Model 1. If $\lambda_1 = \lambda_2 = \dots = \lambda_n = \dots = \lambda$ then the process $N = \{N(t) : t \geq 0\}$ is a well known Poisson process (Kingman (1993)).

If the data collecting process is observed only for the population of finite size n we assume that the numbers $\lambda_1,...\lambda_n,....$ determining the homogeneous pure birth process are distinct and positive numbers and $\lambda_{n+1} = \lambda_{n+2} = ... = ... = 0$.

In this case the sequence τ_k, $k = 1,2,...,n$, $n \geq 1$, forms a limited stream of the respondents' arrivals and is obtained by a superposition (summing) of the streams of n respondents where each respondent generates a stream consisting of exactly one questionnaire. The particular cases of models for the population of the finite size are given below.

Models for the population of finite size

Model 2. If $\lambda_1,...\lambda_n$ are distinct and positive numbers and $\lambda_{n+1} = \lambda_{n+2} = ... = ... = 0$ then we obtain a general homogeneous pure death process.(Getka-Wilczyńska, 2009).

This model satisfies the following conditions:

1. N has the Markov property
2. $\Pr(N(t+h) - N(t) = 1 | N(t) = k-1) = \lambda_k h + o(h), \quad k \geq 1,\ h \to 0+$,
3. $\Pr(N(t+h) - N(t) = 0 | N(t) = k-1) = 1 - \lambda_k h + o(h), \quad k \geq 1$,
4. If $\Pr(N(t) = n)$ then the Internet survey ends and $\lambda_{n+1} = 0$

For this model the system of Kolmogorov differential equations satisfied by $P_n(t) = \Pr(N(t) = n)$ for $t \geq 0$ is given by

$P_1'(t) = -\lambda_1 P_1(t)$,

$P_k'(t) = \lambda_{k-1} P_{k-1}(t) - \lambda_k P_k(t)$, $k = 2,3,...n$,

$P_{n+1}'(t) = \lambda_n P_n(t)$

with boundary conditions $P_1(0) = 1, \quad P_k(0) = 0, k > 1$.

The solution is given by the formulae

$$P_{n+1}(t) = 1 - \lambda_1 \lambda_2 \cdots \lambda_n \sum_{k=1}^{n} \frac{e^{-\lambda_k t}}{\lambda_k w'(-\lambda_k)},$$

where $w(x) = (x + \lambda_0)(x + \lambda_1)...(x + \lambda_n)$.

Model 3. $\lambda_k = \lambda$ for $0 \leq k \leq n$ and $\lambda_k = 0$ for $k > n$. The series $\sum_{k=0}^{\infty} \frac{1}{\lambda_k} = \infty$ is divergent, if any $\lambda_k = 0$. Then $\sum_{k=1}^{\infty} P_k(t) = 1$ and the system of Kolmogorov differential equations is given by

$P_0'(t) = -\lambda P_0(t)$

$P_k'(t) = -\lambda P_k(t) + \lambda P_{k-1}(t)$ for $1 \leq k \leq n$,

$P'_{n+1}(t) = \lambda P_n(t)$ for $k = n+1$ with boundary conditions $P_0(0) = 1, \quad P_n(0) = 0, n > 0$.

The solution is given by the equations

$$P_0(t) = e^{-\lambda t}, P_1(t) = \lambda t e^{-\lambda t}, \ P_2(t) = e^{-\lambda t}\frac{(\lambda t)^2}{2!} \ldots.., P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{k!},$$

$$P_{n+1}(t) = 1 - \sum_{k=0}^{n} e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

Model 4. $\lambda_k = (n+1-k)\lambda$ for $0 \le k \le n$ and $\lambda_{n+1} = 0$

The system of Kolmogorov differential equations is given by

$P'_0(t) = -(n+1)\lambda P_0(t)$

$P'_k(t) = -(n+1-k)\lambda P_k(t) + (n-k+2)\lambda P_{k-1}(t)$ for $1 \le k \le n$,

$P'_{n+1}(t) = \lambda P_n(t)$ for $k = n+1$ with boundary conditions $P_0(0) = 1, \quad P_n(0) = 0, n > 0$.

The solution is given by equations

$P_0(t) = e^{-\lambda(n+1)t}, \ P_1(t) = (n+1)te^{-\lambda nt}\left(1 - e^{-\lambda t}\right), \ldots,$

$P_n(t) = (n+1)e^{-\lambda t}\left(1 - e^{-\lambda t}\right)^n,$

$P_{n+1}(t) = \left(1 - e^{-\lambda t}\right)^{n+1}.$

Properties of model 4

Let $\underline{X} = (X_1, \ldots, X_n)^T$, $n \ge 1$ be a sample of independent and identically random variables selected from the population having an exponential life distribution $F(t) = 1 - e^{-\lambda t}$. $X_k, k = 1,2,\ldots,n$, $n \ge 1$ is interpreted as the waiting time of an arrival (or a response) of u_k th respondent belonging to the population of the size n, $n \ge 1$ and $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ denotes the random vector of order statistics of the sample $\underline{X}$.

In the Internet survey we observe the successive arrivals of respondents at the moments of recording the questionnaires on the server

$$\tau_1 \le \tau_2 \le \ldots \le \tau_n, \ \tau_0 = 0,$$

where $\tau_1 = X_{1:n}, \tau_2 = X_{2:n}, \ldots, \tau_n = X_{n:n}$ and

$T_1 = X_{1:n} \sim Exp(n\lambda)$, $T_2 = X_{2:n} - X_{1:n} \sim Exp((n-1)\lambda)$,…, $T_n = X_{n:n} - X_{(n-1):n} \sim Exp(\lambda)$,

$T_{(0)} = X_{(0)} \equiv 0$.

Properties of the sequences $(T_1,...,T_n)$ and $(\tau_1,...,\tau_n)$ (e.g. Barlow, Proschan (1975), Feller, v. II, (1978)).

1. $P(T_k \le t) = F_{(n-k+1)\lambda}(t) = 1 - e^{-(n-k+1)\lambda t}$ for $k = 1,2,...,n$, $T_1,...,T_k$ are independent random variables, $E\,T_k = \dfrac{1}{(n-k+1)\lambda}$,

$Var(T_k) = \dfrac{1}{[(n-k+1)\lambda]^2}$,

$$f_{(T_1,T_2,...,T_n)}(t_1,...,t_n) = n!\prod_{i=1}^{n}\lambda\, e^{-\lambda(t_1+...+t_n)} = \prod_{i=1}^{n}(n-i+1)\lambda\, e^{-\lambda(n-i+1)t_i}.$$

2. $E(\tau_k) = E(X_{(k)}) = E\sum_{i=1}^{k}\left(X_{i:n} - X_{(i-1):n}\right) = \sum_{i=1}^{k}E\left(X_{i:n} - X_{(i-1):n}\right) = \sum_{i=1}^{k}\dfrac{1}{\lambda(n-i+1)}$,

$$Var(\tau_k) = Var(X_{k:n}) = Var\sum_{i=1}^{k}\left(X_{i:n} - X_{(i-1):n}\right) = \sum_{i=1}^{k}Var\left(X_{i:n} - X_{(i-1):n}\right)$$

$$= \sum_{i=1}^{k}\frac{1}{[\lambda(n-i+1)]^2},$$

$$f_{\tau_k}(t) = \frac{n!}{(k-1)!(n-k)!}\lambda\left(1-e^{-\lambda t}\right)^{k-1} e^{-\lambda(n-k+1)t}.$$

Interpretation: Because for fixed $n \ge 1$ the size of the Internet sample at the moment $t \ge 0$ is given by

$$N(t) = card\,\{1 \le k \le n: \tau_k \in [0,t]\} = \max\{1 \le k \le n : \tau_k \le t\} = \sum_{k=1}^{n} \mathrm{I}_{[0,t]}(\tau_k),$$

and the value of the random variable $N(t), t \ge 0$ equals the number of arrivals up to the time $t \ge 0$, the process $\{N(t),\ t \ge 0\}$ can be described in the following way. Each respondent fills in only one questionnaire with the probability 1 in the interval $[0,T]$, independently from the others. The probability of filling in the questionnaire by certain respondent in the interval $[0,t] \subset [0,T]$ is equal to $P(X_k \le t) = 1 - e^{-\lambda t}$. In this way each respondent generates a stream consisting of only one questionnaire. The summary stream obtained by summing these streams forms a bound Bernoulli stream and consists of a finite number of events.

3. Since X_k, $k = 1,2,...,n$, $n \geq 1$, are independent random variables with exponential distribution $F(t) = 1 - e^{-\lambda t}$, $\lambda > 0, t \geq 0$

then $\Pr(N(t) = k) = \Pr(\tau_n \leq t < \tau_{n+1}) = \binom{n}{k}\left(1 - e^{-\lambda t}\right)^k \left(e^{-\lambda t}\right)^{n-k}$.

Hence, the number of respondents' arrivals up to the time t has binomial distribution with parameters n and $1 - \exp(-\lambda t)$, $N(t) \sim Bin(n, 1 - \exp(-\lambda t))$.

4. The process $\{N\ (t) : t \geq 0\}$ has no independent increments and it is a Markov process with transition probability: for $0 \leq t_1 \leq t_2 \leq .... \leq t_k$

$$\Pr\{N(t_k) - N(t_{k-1}) = k | N(t_1), N(t_2), ..., N(t_{k-1})\} =$$

$$= \binom{n - N(t_{k-1})}{k}(1 - \exp(-\lambda(t_k - t_{k-1})))^k \exp(-\lambda(t_k - t_{k-1}))^{n - N(t_{k-1}) - k}$$

5. The finite dimensional distribution of the process $\{N(t) : t \geq 0\}$ is equal to

$$\Pr(N(t_1) = n_1, N(t_2) = n_2, ..., N\ (t_k) = n_k) = \binom{n}{n_1}\left(1 - e^{-\lambda t_1}\right)^{n_1} e^{-\lambda(n - n_1)t_1} \times$$

$$\times \binom{n - n_1}{n_2 - n_1}\left(1 - e^{-\lambda(t_2 - t_1)}\right)^{n_2 - n_1} e^{-\lambda(n - n_2)(t_2 - t_1)} \times ... \times$$

$$\binom{n - n_{k-1}}{n_k - n_{k-1}}\left(1 - e^{-\lambda(t_k - t_{k-1})}\right)^{n_k - n_{k-1}} e^{-\lambda(n - n_k)(t_k - t_{k-1})}$$

with the covariance function

$$Cov(N(t), N(t+s)) = n \exp(-\lambda(t+s))(1 - \exp(-\lambda t)).$$

4. Conclusion

In this paper some stochastic models to the interpretation and study of Internet mediated research (survey) are proposed. Firstly, we assume that the random sequence of the moments of questionnaires record observed in the Internet data collection process is treated as the stream of events. Next, the size of the Internet sample is described as a counting process built on the points of this sequence. It is a homogeneous pure birth process and there are different versions of it – the Poisson process and the cut models for the population of finite size. The pure birth process and the Poisson process to study the countable population are proposed, the cut models of these processes to study the population of finite size.

Acknowledgments

The research was partially supported by the Ministry of Science and Higher Education programs 03/S/0074/08 and 03/E/0013/09, Warsaw School of Economics.

REFERENCES

BARLOW, R.E, PROSCHAN F. (1975). Statistical theory of reliability and life testing. Holt, Rinehart and Winston Inc., New York 1975.

CALLEGARRO M., DISOGRA CH. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, 72, 5, Special Issue, 1008–1032.

COUPER M.P., MILLER P.V. (2008). Web survey methods. *Public Opinion Quarterly*, 72, 5, Special Issue, 831–835.

FELLER W. (1977, 1978). An introduction to probability theory and its applications (in Polish). Vol. I, II, Polish Scientific Publisher, Warsaw.

GALESIC M. (2006). Dropout on the Web: effects of interest and burden experiences during an online survey. *Journal of Official Statistics*, 22, 313–328.

GETKA-WILCZYŃSKA E. (2009). Mathematical modeling of the Internet surve*y.* Engineering the Computer Science and IT ed. by Safeeullah Soomro, In-Teh, Croatia.

http://www.sciyo.com/articles/show/title/mathematical-modeling-of-the-internet-survey.

GNEDENKO B. W., BIELAJEV A., SOLOWIEV A.D. (1968). Mathematical methods in reliability theory (in Polish). Technical –Scientific Publisher, Warsaw.

KINGMAN J. F.C. (1993). Poisson processes. Clarendon Press Oxford.

LEE S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web Surveys. *Journal of Official Statistics*, 22, 2, 329–349.

LEE S., VALIANT R. (2009). Estimation for volunteer panel Web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods and Research*. Vol.37, No 3, February, 319–343.

RESNICK S.I. (1998). Adventure in stochastic processes. Birkhäuser Boston.

ROSENBAUM P.R., RUBIN D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometricka*, 70, 41–55.

SÄRNDAL C.E., SWENSSON B., WRETMAN J. (1992). Model assisted Survey Sampling. Springer-Verlag New York Inc., New York.

SCHONLAU M., van SOEST A., KAPTEYN A., COUPER M. (2009). Selection bias in Web surveys and the use of propensity scores. *Sociological methods and research*, Vol. 37, No. 3, 291–318.

TILLÈ, Y. (2006). Sampling algorithms. Springer Science-Business Media, Inc., New York.

VEHOVAR, V. (2007). Workshop on Internet Survey Methodology. Ullehammer, September 2007.

http://vv.ris.org, www.WebSM.org

http://www.aapor.org