Statistics Belgium

Working Paper

La Direction générale Statistique et Information économique propose des informations statistiques impartiale. Les informations sont diffusées conformément à la loi, notamment pour ce qui concerne leur confidentialité.

Nous classons les statistiques en huit domaines :

- Généralités
- Territoire et environnement
- Population
- Société
- Économie et finances
- Agriculture et activités assimilées
- Industrie
- Services, commerce et transports

Éditeur responsable : N. DEMEESTER
© 2007, SPF ECONOMIE
DIRECTION GENERALE STATISTIQUE ET INFORMATION ECONOMIQUE | éditeur
B-1000 Bruxelles – 44 rue de Louvain

Statistics Belgium
Working Paper n° 12

An incremental 2-stage sampling plan for a Flemish hepatitis prevalence study: accumulation of respondents over successive waves

Peter Slock, *Camille Vanderhoeft* and *Sophie Quoilin*

Table of contents

Table of contents

An incremental 2-stage sampling plan for a Flemish hepatitis prevalence study: accumulation of respondents over successive waves

An incremental 2-stage sampling plan for a Flemish hepatitis prevalence study: accumulation of respondents over successive waves

Peter Slock, *Camille Vanderhoeft* and *Sophie Quoilin*

Abstract

This paper describes in some detail a methodology for sample survey design, which can significantly improve the quality of the survey process. The proposed sampling design uses a concept of waves (spread over time) for contacting selected individuals, combined with a certain grouping of the sampling units. The purpose is to obtain an initial sample and a set of respondents that are highly representative of the study population, according to various criteria. Another asset of the described methodology is the close control that is obtained over the number of respondents already at the sampling design stage. The paper presents the mathematical background for the proposed approach and extensively illustrates its application to the hepatitis prevalence study conducted recently in the Flemish population.

1 The Flemish Hepatitis Prevalence Study (HPS)

1.1 *Background and survey objectives*

The *Hepatitis Prevalence Study* (HPS) is a survey conducted by the Belgian *Institute of Public Health* (IPH). This is a scientific institute of the State (Royal Decree of 6th March 1968) which also has the "legal personality" in order to facilitate the execution of contracts for third parties. The IPH consists of three departments: Microbiology, Pharmaco-Bromatology and Epidemiology-Toxicology. More specifically the tasks of the Epidemiology-Toxicology unit, which organises the HPS, are:

- to gain insight into the population's health status and its determinants;
- to monitor trends in health status and to organise surveillance systems for a series of diseases and health determinants;
- to promote quality of care through health service research;
- to provide information to public health stakeholders for their decision making processes;
- to co-ordinate health information in Belgium.

The topic of the above mentioned survey is hepatitis prevalence in the Flemish region.

Viral hepatitis includes several distinct infections affecting the liver. Most frequent viruses are hepatitis A, hepatitis B and hepatitis C. They have similar clinical presentation but differ in some aspects as occurrence, mode of transmission or clinical evolution.

It is important to have complete information over the trends in the prevalence of viral hepatitis because a relatively high number of persons can be exposed to hepatitis A, B or C and because morbidity and mortality associated with these diseases can be serious (especially for hepatitis B and C). The implementation and the evaluation of efficient public health measures depend largely on a good knowledge of the epidemiology of a disease.

Different methods can be used to follow the epidemiology of a disease. The hepatitis prevalence study (HPS) is one of them. The HPS is characterised by the use of saliva to identify antibody. This method of identification of the diseases has permitted to perform a study in the general population: a sample of individuals can be drawn and a test set can be sent by mail to selected individuals, who return the set after using by mail too.

Initially, the whole Flemish population (i.e. all Dutch speaking people in Belgium) was targeted, but because of the difficulties to identify Dutch speaking people in the Brussels Capital Region (which is bilingual) the study was limited to the 5 Flemish provinces.

1.2 *Main objectives of the sampling design*

IPH specified an initial study design based on a certain sampling scheme. Based on the known prevalence of the diseases in the Flemish region (Beutels *et al*, 1997), the sample size has been calculated with *Epi-Info* 6.4d[a]. A minimum of 1400 persons has to be included in the study. There is no real upper limit for the number of participants in the study, but IPH preferred the number of participants not to exceed 2000 because of practical constraints such as availability and cost of test material and doctors.

Another constraint was time: the study had to take place within a 4 months period starting from the contacting of the first individuals. Notice that in the sequel of this text, *individual* and *person* are synonyms.

The main objectives to be considered when designing the sample are:

- to keep close *control over the number of respondents* (i.e. the number of people participating in the hepatitis study);
- to obtain a (respondent) sample which is *representative* of the study population.

Here, *representative* means that the main characteristics of the sampling units should be distributed more or less in the same way in the (respondent) sample and in the study population. The main characteristics of the individuals for the HPS are age (category), sex and province of their domicile (as well as municipality and statistical letter of domicile). Of course, if the distributions of these variables differ (slightly) between the study population and the respondent sample, response modelling and/or calibration can be applied to correct for this deficiency afterwards. However, when a large bias is introduced in the results because of certain groups of individuals (e.g.

[a] Epi-info is a statistical software tool for epidemiology, developed by the US "Centre for Diseases Control" in Atlanta and freely downloadable from their website http://www.cdc.gov/epiinfo/.

females aged between 15 and 24) being underrepresented in the study, calibration cannot always completely remove bias. Thus, whatever methods are used to tackle non-sampling errors, it is always advisable to try to control for the number of respondents and representativity (in the strict mathematical sense) of the respondent sample.

Statistics Belgium, which has easy access to the majority of variables in the Belgian National Population Register[b] (NPR), started from the NPR for building the sampling frame from which it then drew the sample for HPS. The NPR contains all variables cited above which are needed with respect to representativity. The sampling frame contains some 6 million records, each corresponding to 1 individual from the NPR.

There are several ways to obtain representativity for a sample. First, the criterion or criteria according to which representativity is desired can be used as stratification criteria. As it is quite important to obtain results that are sufficiently accurate per province[c], representativity per province was considered to be of primary importance. Hence, the 5 provinces were identified as strata. Of course, the other criteria of representativity (statistical letter[d], age and sex) could also be used as stratification criteria.

Proportional allocation was used for allocating the global sample size to the different strata (= provinces), which allowed obtaining representative samples in a strict mathematical sense for what concerns its spread over the Flemish provinces.

When several stratification criteria are involved – say *A*, *B*, *C* and *D*[e] – problems may arise if the complete cross-classification $A \times B \times C \times D$ is considered: small or empty strata, too many strata … and statistical consequences in estimation. Moreover, non-response can heavily distort the intended sample design. Therefore, the IPH and Statbel adopted the following approach: only one criterion *A* is used for classical stratification. Within each *A*-stratum the remaining criteria (*B, C* and *D*) are used to define the PSUs (primary sampling units, which here correspond to collections of individuals, each of which can be considered as a subpopulation), rather then strata. The size of the PSUs differs largely over the sampling frame.

This approach is justified because no precision requirements exist for criteria *B, C* and *D*: it is even not required that each combination of these 3 properties is present in the

[b] The NPR is a government database which identifies in a unique way all people (individuals as well as the households they belong to) who have their domicile in Belgium (also foreign people, as long as they live in Belgium) and contains many interesting characteristics about these people and households (such as their address, names etc.). For more information on the NPR, http://www.rijksregister.fgov.be/index.htm (available in the 3 national languages: Dutch, French and German).

[c] The 5 Flemish provinces correspond to the NUTS-2 areas in the Flemish region.

[d] A statistical letter is part of a Belgian municipality. The statistical letters correspond more or less to the old municipalities (which existed till the merger of Belgian municipalities in the seventies). There are 2644 statistical letters in the whole territory of *Belgium*. Notice that the number of statistical letters in a province or region is *not* proportional to its number of inhabitants.

[e] *A* corresponds to the provinces (see strata defined above), *B, C* and *D* correspond respectively to statistical letter, age category and sex.

sample, whereas it is important that each province (see criterion *A*, used for stratification) is represented (to a sufficient amount) in the sample.

1.3 *History of the sampling design for HPS*

IPH initially proposed (June 2003) to draw a *systematic sample* of individuals, from a frame sorted by age, sex, municipality and statistical sector[f]. The proposed sample size was 20000, to be allocated proportionally to the 5 Flemish provinces (NUTS-2 regions, or *strata*). The sample should then be split into 4000 *groups*, containing 5 individuals each. In a first wave, an invitation for participation in the survey would be sent to 4000 selected individuals, one from each group. Those 4000 individuals would have 3 weeks to respond: hopefully enough among them would respond positively (meaning that they were willing to participate). The response rate was expected to be about 10%.

After 3 weeks, groups with a recorded positive response would have to be "inactivated". Next, a second mailing would again be sent to one other selected person in each active (i.e. not inactivated) group. Again, after another 3 weeks of response recording, responding groups would be inactivated, and a third mailing would be organised. This process would continue until the fifth (planned) mailing was sent and response was recorded. The whole process would last for 15 weeks. The objective was to find at least 1500 (positive) respondents, and a final respondent sample that is representative with respect to age, sex, statistical sector and municipality.

For the above procedure, the expected number of respondents can easily be calculated with the formula $m\left(1-(1-p)^{W}\right)$: with $m = 4000$ groups, expected response rate $p =$ 10% and $W = 5$ *waves* (i.e. mailings), the expected number of respondents turns out to be about 1638, and an expected number of about 16380 individuals will be invited to participate. The desired number $R = 1500$ of respondents would be reached! Moreover, it is easy to see that the resulting respondent sample is very likely to be representative.

Statistics Belgium started analysing the problem, and together with IPH it was soon decided to consider the following initial sampling procedure:

- from a frame of individuals, construct the strata;
- allocate a total sample size proportionally to strata;
- within strata, construct PSUs according to individual characteristics (domicile, age and sex);
- draw PSUs systematically (with replacement), within each stratum;
- within sampled PSUs, draw a number of individuals with SRS (without replacement);
- within sampled PSUs, the selected individuals are distributed over one or more groups (of equal size).

[f] A statistical sector is a small part of a statistical letter and hence a (very) small part of a Belgian municipality.

This basically 2-stage sampling design is self-weighting, and, by the law of large numbers, the resulting initial sample is representative with respect to the characteristics used to construct strata and PSUs.

IPH still wanted to contact sampled individuals in successive waves, but no agreement about the parameters (the number W of waves and the number C of persons to be invited in each wave and group, among other parameters) could be reached in that early phase (before summer 2003) of the discussions. Statistics Belgium argued that many sets of parameters could be considered and evaluated in terms of the expected final number of respondents. Actually, the mathematics of the procedure were studied gradually, prototype simulation programs were developed and results of simulations were presented to and discussed with IPH. More specifically, we studied the distribution of the final number of respondents and noticed (in September 2003) that we could have a lot of control over the expected number of respondents (after each wave).

At that time (September 2003), after studying and discussing several alternative sets of parameters, it was decided to draw $n = 24000$ individuals in $m = 3000$ groups of size $G = 8$ each, planning $W = 4$ waves, and inviting $C = 2$ individuals per wave and per group (as long as a group remains active). Simulations and exact calculations showed that, with a hypothetical $p = 12\%$, the expected number of respondents would be about 2151 (after 4 waves, and about 1800 after 3 waves…).

More details about the calculations and simulations that have led to the final sampling design for HPS are discussed in section 4.

The present paper is largely devoted to the mathematical framework of the procedure that has thus far been outlined in general terms only. We also discuss the simulation procedures that have been designed (and implemented), based on that framework, and how these simulations (and exact calculations) can be used in the design phase of a sample survey. The mathematical details are given in section 3, and in Appendices A, B and C. Simulation results are summarised in Appendix E and discussed in section 4.

We believe that these techniques can be very useful for many surveys, especially those where non-response is really an important issue. Traditionally the problem of non-response has been tackled in several different *ad hoc* ways. This will be discussed briefly for some household and individual surveys at Statistics Belgium in the next section.

2 Dealing with non-response at Statbel

There are basically 4 ways of dealing with non-response at the moment of working out the sampling design of a survey: *grossing up* the desired number of respondents (dividing it by the response rate), *quota* sampling (surveying no more than a fixed

number of respondents per subgroup[g]), *substitutes* (sometimes called "replacement units") and, finally, *waves* of contacting the sampling units. All these methods take non-response already into account at the sampling stage.

Let's now briefly describe the disadvantages and advantages of the 4 proposed solutions.

2.1 *Grossing up the desired sample size*

This solution starts from the *desired* number of units participating in the survey. As the response rate for a survey is always smaller than one, the division of the desired number of respondents by the expected response rate yields a larger number of units to be selected into the sample (hence the term "grossing up"). The latter is the number of units required in the initial sample in order to arrive at the specified value for the desired number of respondents, *assuming* that the real (i.e. observed) response rate corresponds to the estimate used for grossing-up.

Of course, small variations in the estimate for the response rate can lead to huge variations in the obtained estimate of the required number of sampling units, and this estimate only reflects one's *expectations*. The expected response rate may turn out to be quite different from the observed rate for the survey being dealt with.

The big difficulty with grossing-up consists of finding a good estimate for the response rate beforehand. In case of a survey which is repeated in time, one can estimate the response rate from the observed response rate during one or several previous survey editions, or from the observed response rate from a corresponding pilot survey, but when such surveys are not available from the past (e.g. in case of a new or reorganised survey), estimating the response rate for the forthcoming survey is mostly a pure guess.

2.2 *The case of HBS: eliminating positive response by quota sampling*

Quota sampling is a non-probabilistic, *ad hoc* sampling method having several disadvantages. First of all, its non-probabilistic nature causes the traditional theory of survey estimation (extrapolation and variance estimation) not to be applicable as such. Furthermore, sampling units which are prepared to participate in the survey can be refused irreversibly as soon as the quota has been reached in a given stratum or "quota group". If it appears afterwards that, because of dropout, the final number of participants is (considerably) lower than the quota, the refused sampling units cannot be re-entered into the sample. Hence, this method often results in a number of participants that might be (far) below the number of participants that can be realised with the available budget.

The only advantage of quota sampling is that the budget can be closely controlled per "quota group", as no more participants than a pre-specified number (equal to the

[g] Quota sampling often pre-specifies subgroups within the sample and then only allows a certain number of respondents per subgroup, thus ensuring that within each subgroup the number of respondents (and corresponding cost) does not exceed a certain threshold.

specified "quota" limit) can occur within each group, thus simplifying the financial administration for the survey.

Because of its various disadvantages, quota sampling is never applied in its pure form at Statistics Belgium, but only after some probabilistic sampling steps took place. Quota sampling techniques are used in the Household Budget Survey (HBS) at Statistics Belgium: randomly selected individuals are invited for participation, but positive responders can be removed if the corresponding quota group is "complete". Quota sampling has been introduced in the HBS for two reasons: it is intended to improve the representativity of the final respondent sample, and it is preferable (merely from an organisational point of view) to assign he same number of respondents to each interviewer.

2.3 *The case of LFS: substitute households*

The use of substitutes is another *ad hoc* sampling method for dealing with non-response yet at the sampling stage of a survey. Its main disadvantage is a huge danger of bias in the survey results, because it's very tempting for interviewers to replace units that are hard to contact or survey by other units (among the "substitutes") that are easier to contact or survey. In fact, the final composition of the sample (namely those units which are effectively contacted for survey participation) depends partly on how the sampling units are contacted, as it is explained below for the implementation of substitutes in surveys at Statistics Belgium.

Substitutes are often used in social statistics, and mainly in surveys dealing with households or individuals. At Statistics Belgium, the Labour Force Survey (LFS) uses replacement households; each group of households, which has been assigned to a certain interviewer, contains some substitutes. Substitutes are contrasted to the other households in the group, which are called effective households. The substitutes should only be contacted when there are no means any more of contacting and surveying some of the effective households of that group. As exclusively the interviewer for that group does the contacting of the household, the composition of the respondent sample for a certain group depends on that interviewer. This can introduce some bias into the estimates.

Because of the danger with substitutes, which should not be underestimated (Slock, 2004 and Dawagne, 2004), the EU-SILC[h] 2003 survey did not adopt substitutes, but operated via a system of waves instead.

2.4 *Contacting sampling units in waves*

Finally, a technique based on the idea of *waves* can be considered for contacting the sampling units. IPH launched this idea and after bilateral discussion Statistics Belgium and IPH decided to adopt the wave concept (as it was introduced yet in section 1.3) for organising the contacting procedure of the hepatitis survey. The main issue with this concept is that the different sampling units are not contacted all at the same time, but in different waves spread over time.

[h] Statistics on Income and Living Conditions, survey enforced by the European Commission (Eurostat).

This way of contacting the sampling units, although pretty new at Statistics Belgium, looks a quite promising method, the principles of which could be applied to other sample surveys as well. Its main asset is that it allows to reach almost for sure the desired number of respondents, without exceeding that number too much, such that the budget foreseen for the survey is not exceeded and the survey costs are hence controlled closely.

The only disadvantage related to this method at present appears to be the longer period which is needed for contacting the sampling units: some 3 or 4 months have to be available for organising the different survey waves.

Statistics Belgium gained some experience with waves during last year's edition of the EU-SILC survey. The survey was then organised via 2 waves. After the 1st wave, the *observed* response rate was calculated per stratum and these response rates were used as the *expected* response rates for the respective strata during the 2nd wave. This yielded a final total number of participants (generally called "respondents") that was situated within the prescribed interval for the desired number of respondents.

There are some similarities between the wave system as used in the HPS and the system of substitutes described in the previous subsection. In both cases, the initial sample is very likely to be larger than the part that will be effectively used. With "effectively used", those units are meant who are actually contacted for survey participation. In the HPS sampling approach, the notion of waves is combined with a specific kind of *groups*, which are constituted from PSUs, where each PSU collects all individuals that have identical values for the criteria of representativity (viz. statistical letter, age category and sex).

In the HPS survey, the *initially contacted individuals* for a certain group (corresponding to the first wave) are "replaced" by an equal number of individuals from the remaining part of the same group in the 2nd wave (if a 2nd wave is organised) if none of the contacted individuals in that group responded positively during the 1st wave. This procedure is repeated for each group during subsequent waves as long as no respondent has been found yet for that group and the total desired number of respondents has not been reached yet. When using waves, the replacement of sampling units by other units is controlled centrally, based on the total number of respondents that far. Moreover, when some other individuals are contacted for a certain group (because the individuals contacted earlier in that group do not participate), the replacement system used in the HPS wave approach assures that the newly contacted individuals have the same characteristics as the initial ones, as they belong to the same group.

When using substitutes instead, and when interviewers can manage their own groups containing both substitute and effective units, there is no guarantee at all that the substitutes have the same characteristics as the effective households that are replaced by them. The system of substitutes used for the Belgian LFS does not control the substitution mechanism in a central way, thus introducing a serious danger of interviewer bias.

Mathematical details underlying the wave concept - as it was tried out for HPS - are described in section 3.

2.5 *Different types of samples*

Four kinds of samples can be distinguished for the hepatitis survey as well as for other surveys that take place via the above wave mechanism. Each of these sample types corresponds to a certain step in the survey process:

1. The *initial sample*: as described above, this sample contains all individuals selected from the national population register (24000 people in case of the recent hepatitis survey); one should stress that this sample merely relates to the *selection* process from the NPR (it does not matter whether these sampled units are effectively contacted afterwards for survey participation, whereas in traditional sample surveys all units that were selected into the sample are contacted).
2. The *contact sample*: that part of the initial sample whose elements are actually contacted during some of the *organised*[i] waves; here, 'contacted' means that they are *invited* by mail to participate in the survey; hence the final sample consists of the above sampled individuals (see *initial* sample) to whom an advance letter was sent (6000 people, all corresponding to the 1st wave, in case of the recent hepatitis survey).
3. The *initial respondent sample*: those units from the contact sample who declared themselves prepared to participate in the survey (some 2000 people, responding during the 1st wave).
4. The *final respondent sample*: those units from the initial respondent sample for whom test results were obtained at IPH, i.e. those people that not only declared their willingness to participate but for whom a sample of saliva[j] was taken by a medical doctor and received by IPH as well. Notice that this set is significantly smaller (viz. some 1800 people) than the previous one (initial respondent sample) because of the occurrence of dropout: some people who were (initially) willing to participate changed their mind or the sample of saliva obtained for that person was not usable or did not arrive at IPH. Dropout has an impact on the final response rate. It can e.g. result in a final respondent sample that is smaller than indicated by the desired number of respondents, even if the initial number of respondents (i.e. the size of the previous set) is larger than that desired number. It can also cause the preliminary elimination of a group from subsequent waves[k], whereas finally no test results were obtained from a person of that group, because that person did not participate in the end or because no test results were received for them at IPH. Unfortunately, it is not possible to predict dropout in a reliable way for surveys such as the HPS, because of various uncontrollable factors that may cause

[i] *Organised* means that the wave took actually place in practice. It can happen however that a certain maximum number of waves W is foreseen but that the procedure of contacting individuals is already stopped after $w<W$ waves ($w=1$ in the case of our survey).
[j] to identify antibody
[k] Because at least 1 person of that group answered positively to the advance letter.

dropout. Therefore, dropout was not taken into account in the modelling when designing the sample.

3 An incremental sampling plan: the SIR procedure

3.1 *Population and sampling parameters*

Consider a population[l] U that consists of N individuals. According to some individual characteristics (such as place of domicile, age, sex…), U is divided into a number M of PSUs U_i with size (number of individuals) N_i ($i = 1, \ldots M$). Notice that $U = \bigcup_i U_i$ and $N = \sum_i N_i$.[m]

It is supposed that the following sampling parameters have been fixed:

- n : the size of the initial sample of individuals to be drawn from U;
- m : the size of the sample of PSUs to be drawn (with replacement);
- W : the number of planned waves;
- G : the number of individuals to be selected (without replacement) from a selected PSU for each draw of that PSU;
- C : the number of individuals from a group of G selected individuals that have to be contacted per wave;
- R : the desired number of respondents;

with the following relationships: $n = mG$, $G = WC$.

3.2 *Sampling and contacting procedure*

We now introduce the following 5-step procedure of sampling, inviting to participate and response recording, called the *SIR procedure*.

Step 1 : *Stage* 1*, i.e. sampling of PSUs*
Systematic PPS sampling[n] is applied to select m times a PSU. The inclusion probability for PSU i equals 1 if the size $N_i \geq a$, and equals N_i / a if $N_i < a$. The *multiplicity* of selecting the i-th PSU is denoted as m_i. Notice that $m_i \geq 0$, and $m = \sum_i m_i$. The (random) number of different selected PSUs is $m' = \sum_i \min(1, m_i)$. The label "PPS" for this sampling

[l] The "population" may be a sub-population, or stratum in case of a priori stratification, within the total study population. Our development in the present section thus applies to each sub-population separately.

[m] $\bigcup_i$ resp. $\sum_i$ always stand for a union resp. a summation over all PSUs i in the population U.

[n] A fractional interval method is applied, with interval (or step) $a = \sum_i N_i / m = N/m$ and with random start $b \in [0; a]$. This guarantees a fixed size sample of m (draws of) PSUs.

method is justified by the property that the expected multiplicity for each PSU *i* is proportional to its size N_i.

Step 2 : *Stage* 2*, i.e. sampling of individuals (SSUs)*
In each selected PSU *i*, i.e. PSU with $m_i \geq 1$, m_i *groups* of individuals have to be drawn. Each group selected from PSU *i* will contain *G* individuals, and the m_i groups in PSU *i* are supposed to be non-overlapping. The latter can be guaranteed by drawing all $m_i G$ individuals at once, and constructing separate groups afterwards.
SRS without replacement is applied to select $m_i G$ individuals. Hence the conditional inclusion probability for an individual *k* in PSU *i* equals $m_i G / N_i$.[o]
Let $\mathcal{G}^{(1)}$ be the set of groups from all selected PSUs. The set $\mathcal{G}^{(1)}$ has size *m*. The *initial sample*, resulting from steps 1 and 2, is simply the union of the *m* groups in $\mathcal{G}^{(1)}$, and has fixed size $n = mG$.

Proposition 1 : The overall *initial inclusion probability*, induced by the 2-stage sampling procedure described in steps 1 and 2, for individual *k* in PSU *i* equals $\pi_{ik}^{(0)} = G/a = n/N$. Notice that the proposed 2-stage sampling design is self-weighting.

Remark 1 : If all *n* selected individuals were invited to participate in the survey, then the corresponding sampling weights (when using the Horvitz-Thompson estimator of a population total) one would use for estimation were $d_{ik}^{(0)} = 1/\pi_{ik}^{(0)} = a/G$.

Step 3 : *Assigning selected individuals to waves*
The *G* selected individuals in each group of a selected PSU *i* are randomly assigned to *W* waves, with the same number *C* of individuals being assigned to each wave ($G = WC$).

Step 4 : *Contacting and observing response in wave* 1
In the first wave, the *C* individuals that are assigned to this wave (step 3) in each group of a selected PSU *i* are invited to participate in the survey. Hence *mC* individuals are invited for participation. A total number x_1 will respond positively. These respondents belong to a number $\dot{x}_1$ of so-called *responding groups*: a group is said to be responding, if at least one of its members has responded positively. Let $X_1 = x_1$ and $\dot{X}_1 = \dot{x}_1$. The

---

[o] It is supposed that $m_i G \leq N_i$ for any PSU *i*. This can be guaranteed before drawing the sample by choosing *G* such that $G \leq N_i \Big/ \left(\left[\frac{N_i}{a} \right] + 1 \right)$ for PSUs for which N_i is not a multiple of *a* and $G \leq N_i \Big/ \left[\frac{N_i}{a} \right] = a$ for PSUs for which N_i is a multiple of *a*. The notation [*x*] stands for the integer part of the real number *x*.

responding groups become *inactive* and are eliminated from the set $\mathcal{G}^{(1)}$. Let $\mathcal{G}^{(2)}$ be the set of groups that are still *active* after wave 1. This set has size $m - \dot{X}_1$.

Stopping rule 1 : The sampling and contacting procedure stops after wave 1 if $X_1 \geq R$ (a number R of desired respondents has been reached[p]), or if $\mathcal{G}^{(2)} = \varnothing$ (there are no more active groups), or if W=1 (only 1 wave was planned). We then set $\mathcal{G}^{(\infty)} = \mathcal{G}^{(2)}$, where $\mathcal{G}^{(\infty)}$ can be interpreted as the set of groups in which no respondent has been found when the procedure is stopped.

Stopping rule 2 : Stopping rule 1 can be simplified, by ignoring R : the sampling and contacting procedure stops after wave 1 if $\mathcal{G}^{(2)} = \varnothing$ or if W=1. This rule will be used in our simulations.

Proposition 2 : If the procedure stops after the first wave, the *initial weight* (used in the Horvitz-Thompson estimator for a total) for any individual k in PSU i is calculated as $d_{ik}^{(1)} = a/C$. The *observed global response rate* would be $X_1/(mC)$ at individual level and $\dot{X}_1/m$ at group level.

Remark 2 : If W=1, then G=C and $d_{ik}^{(1)} = d_{ik}^{(0)}$.

Step 5 : *Inviting selected individuals and recording response in waves w*>1
Now, for any wave w = 2, … W, let X_{w-1} be the total cumulative number of respondents after wave $w-1$, $\dot{X}_{w-1}$ the total cumulative number of responding groups after wave $w-1$, and $\mathcal{G}^{(w)}$ the set of remaining active groups at the start of wave w, with size $m - \dot{X}_{w-1}$.
In wave w ($2 \leq w \leq W$), C individuals in each group belonging to $\mathcal{G}^{(w)}$ and assigned to wave w, are invited to participate in the survey. Hence $\left(m - \dot{X}_{w-1}\right)C$ individuals are now invited for participation. Among these individuals, a total number x_w will respond positively. Those respondents belong to a number $\dot{x}_w$ of responding groups. We then have $X_w = X_{w-1} + x_w$ and $\dot{X}_w = \dot{X}_{w-1} + \dot{x}_w$. We further reduce $\mathcal{G}^{(w)}$ to $\mathcal{G}^{(w+1)}$ by eliminating groups that became responding in wave w. This set $\mathcal{G}^{(w+1)}$ of remaining active groups after wave w has size $m - \dot{X}_w$.

Stopping rule 1 : The sampling and contacting procedure stops after wave w if $X_w \geq R$, or if $\mathcal{G}^{(w+1)} = \varnothing$, or if W=w. We then set $\mathcal{G}^{(\infty)} = \mathcal{G}^{(w+1)}$, where $\mathcal{G}^{(\infty)}$

[p] A safety margin should be added to this threshold R, viz. some 10 or 20% extra respondents to counteract drop-out later on during the survey (supposing that this drop-out is between 10 and 20% of the R initial respondent, see section 4.2).

can be interpreted as the set of groups in which no respondent has been found when the procedure is stopped.

Stopping rule 2 : Stopping rule 1 can be simplified, by ignoring R : the sampling and contacting procedure stops after wave w if $\mathcal{G}^{(w+1)} = \varnothing$ or if W=w. This rule will be used in our simulations.

Notation : Let w_{ij} be the wave after which group ij (group j in PSU i) becomes inactive, with $w_{ij} = \infty$ if the group hasn't become inactive after the wave at which the procedure stops. I.e. $\mathcal{G}^{(\infty)} = \left\{ ij \in \mathcal{G}^{(1)} \middle| w_{ij} = \infty \right\}$.

Let $\tilde{w}_{ij} = \min\left(w_{ij}, w\right)$ if the procedure stops after wave w; $\tilde{w}_{ij}$ is the number of waves in which individuals belonging to group ij have been invited for participation.

Finally, let $\tilde{s}_{ij}$ be the set of individuals in group ij that have been invited for participation in the survey. Notice that $\tilde{s}_{ij}$ has size $\tilde{w}_{ij}C$.

Proposition 3 : If the procedure stops after wave w, the *initial weight* for any individual k in group j of PSU i is calculated as $d_{ijk}^{(w)} = a \Big/ \left(\tilde{w}_{ij}C\right)$. The *observed global response rate* is $X_w \Big/ \sum_{ij \in \mathcal{G}^{(1)}} \tilde{w}_{ij}C$ at individual level and $\dot{X}_w \Big/ \sum_{ij \in \mathcal{G}^{(1)}} \tilde{w}_{ij}$ at group level.

3.3 *Initial weights*

The following property shows that the proposed initial weights have the desired property of summing up to the population size N. This is an argument in favour of the proposed initial weights.

Property 1 : If the sampling and contacting procedure stops after wave w, then $\sum_{ij \in \mathcal{G}^{(1)}} \sum_{k \in \tilde{s}_{ij}} d_{ijk}^{(w)} = N$.

Proof

Since $a = \frac{N}{m}$, we have:

$$\sum_{ij \in \mathcal{G}^{(1)}} \sum_{k \in \tilde{s}_{ij}} d_{ijk}^{(w)} = \sum_{ij \in \mathcal{G}^{(1)}} \sum_{k \in \tilde{s}_{ij}} \frac{a}{\tilde{w}_{ij}C} = \frac{N}{m} \sum_{ij \in \mathcal{G}^{(1)}} \frac{1}{\tilde{w}_{ij}C} \sum_{k \in \tilde{s}_{ij}} 1 = \frac{N}{m} \sum_{ij \in \mathcal{G}^{(1)}} 1 = \frac{N}{m} m = N .$$

□

Remark 3 : The notation introduced before Proposition 3 can be used also in case of stopping the procedure after wave w=1. Then $w_{ij} = 1$ or $w_{ij} = \infty$, depending on whether at least one respondent has been found in group ij or not, and $\tilde{w}_{ij} = 1$ for

all *ij*. It follows that $d_{ijk}^{(w)} = a/\left(\tilde{w}_{ij}C\right)$ in Proposition 3 indeed reduces to $d_{ijk}^{(1)} = a/C$ as in Proposition 2.

Remark 4 : If the procedure stops after wave 1, then one could artificially construct $\pi_{ijk}^{(1)} = 1/d_{ijk}^{(1)} = C/a$. Interpretation of these quantities as inclusion probabilities is equivalent to acting as if only the *mC* individuals invited in wave 1 had been selected, and all other sampled individuals were ignored. The situation is even more cumbersome if the procedure stops in wave *w*>1, because then the quantities $\pi_{ijk}^{(w)} = 1/d_{ijk}^{(w)} = \tilde{w}_{ij}C/a$ could be interpreted as inclusion probabilities if only $\tilde{w}_{ij}C$ individuals in group *ij* had been sampled and invited for participation, while another $\left(W - \tilde{w}_{ij}\right)C$ sampled but not invited individuals in group *ij* had been ignored.

Remark 5 : It follows from Remark 4 that the "initial weights" $d_{ijk}^{(w)}$ aren't sampling weights in the usual sense. However, the initial weights for respondents can as usual be adjusted for non-response and they can also be calibrated using auxiliary information. The resulting modified weights can then be used for estimation.

Remark 6 : Let $\tilde{s}_{ij}^{r}$ be the set of respondents in group *ij*; its size r_{ij} is at most *C*. Let $\tilde{s}_{i}^{r}$ be the set of respondents in PSU *i*; its size r_i is at most $m_i C$. Since PSUs are homogeneous with respect to the criteria defining PSUs, one could suppose that response homogeneity groups (RHGs) coincide with PSUs, and adjust the initial weights in PSU *i* by a common factor $\phi_i = \left(r_i / \sum_{j \subset i} \tilde{w}_{ij} C \right)^{-1}$ for all groups *ij*. Hence, $\phi_i d_{ijk}^{(w)}$ could be proposed as weights after adjustment for non-response. Notice that r_i has to be strictly positive for each selected PSU *i* for this correction to work.

Remark 7 : In Appendix D the SIR procedure is schematically outlined for a small hypothetical example.

3.4 *SIR simulation*

The 5-step procedure of (simulating) sampling, inviting to participate and registration of response, as described mathematically in section 3.2, will be called *SIR simulation* hereafter.

Given is a frame of individuals (ultimate sampling units), with information about place of domicile, age and sex.

The following parameters have to be chosen:

- *n*, *W* and *C*,
- *p* = global response rate (also the response probability of an individual that was selected and contacted),

- stratification variable(s) and allocation rule,
- PSU variable(s).

It should be noticed that n is here the overall initial sample size, i.e. the number of individuals to be selected from all strata together.

Compute:
- $G = WC$,
- PSU classification variable(s), if needed (e.g. recode age into age classes).

Starting from the frame, the following preliminary actions are taken before the SIR simulation actually starts:

- Frame units are aggregated by (stratification and) PSU variables[q]; PSU sizes N_i are computed and stored in a so-called PSU-table.
- PSUs are further aggregated by stratification variable(s); stratum sizes[r] (N) are computed and stored in a so-called STRATUM-table.
- Allocation of n to the strata[s] (in our programs a choice can be made between N-proportional and $\sqrt{N}$ -proportional allocation); the n in each stratum is rounded to the closest integer multiple of G.
- Since sampling is essentially independent between strata, and PSUs are drawn in the first stage, the number and total size of PSUs per stratum is calculated and stored in the PSU-table.
- PSUs with $N_i < G$ are identified. No further action is taken with respect to these PSUs, and it is hoped that such PSUs will not be selected[t].
- The number of PSU draws within each stratum is computed: $m = n/G$.

After a randomisation seed is set[u], SIR simulation starts. This procedure is discussed in steps 1-5 in section 3.2; stopping rule 2 is used throughout our simulations. So the details need not be repeated here and it suffices in this section to refer to a number of interesting output tables, showing information that can be useful for evaluation of a SIR simulation.

Table S1 :	m and m' (per stratum and total).
Tables S2-S4 :	Marginal and joint distributions of PSU variables among all individuals in sampled PSUs.
Table S5 :	Verification of sample sizes n.

[q] As 'statistical letter' (one of the PSU variables) is part of a certain province (= the stratification variable), it follows that the stratification variable is itself a PSU variable.
[r] As in the preceding section, no index for stratification is used in our notations.
[s] From now on, i.e. after the allocation is computed, n denotes the initial sample size per stratum.
[t] As a future improvement of the sampling design, the selection of such a 'small' PSU has to be *avoided*, by merging it with a 'neighbouring' PSU yet before drawing the sample, such that the merged PSU has size $N_i^* \geq G$. Even more, each PSU i (or *merged* PSU) should be of size $N_i \geq m_i G$, but its multiplicity of selection i is only known after the selection of the PSUs. However, if group size G has been chosen appropriately small (as indicated in footnote o), all of the (non-merged) PSUs will satisfy $N_i \geq m_i G$.
[u] Automatically, but random, by the system, or fixed be the user.

Tables S6-S8 :	Marginal and joint distributions of PSU variables among sampled individuals.
Tables IR1-IR*W* :	Invitation and response in waves 1 to *W*.
Tables CR1-CR*W* :	Cumulative response after waves 1 to *W*.
Tables R1-R3 :	Marginal and joint distributions of PSU variables among simulated respondents.

All these tables are too voluminous to be included in this text but can be readily obtained from the authors. Tables of distributions of PSU variables in the sampling frame are available as well, and comparison of these with the sets of tables S6-S8 and R1-R3 allows evaluating representativity of initial sample and simulated respondent sample with respect to PSU variables.

3.5 *Global simulation of inviting to participate and response recording in waves*

It has to be stressed that this simulation does not include initial sampling. Actually, it doesn't start from a real, or even fictitious, frame. One could imagine that the process starts from a set of (*m*) sampled groups of size *G* each, but even this is not explicit in the proposed global simulation process. Nevertheless, it is a very useful and fast simulation technique, as it allows finding an estimate of the expected number of respondents quickly and accurately, for many different sets of input parameters, together with estimates for a number of other characteristics (such as the variance) of the distribution of the number of respondents.

Choose the following parameters:
- *m*, *W*, *C* and *p*,
- *S* = number of simulations.

Compute:
- $G = WC$ and $n = mG$.
- $\dot{p} = 1-(1-p)^C$ = the probability that among a set of *C* invited individuals at least 1 respondent is found. It's also the probability that a group, which is still active, responds in a particular wave. (It reduces to *p* if *C*=1.)

After a randomisation seed is set, the following steps (i) – (iv) are repeated *S* times:

(i) Set $\dot{X}_0 = 0$ and $X_0 = 0$. Set $w = 1$.

(ii) Generate $\dot{x}_w \sim B\left(m - \dot{X}_{w-1}, \dot{p}\right)$, i.e. the number of responding groups in wave *w*, from the remaining $m - \dot{X}_{w-1}$ active groups.
Generate $\dot{x}_w$ times a number *y* of additional individual respondents within a responding group of size *C*. (See Property A.4 in Appendix A and *RNG* B.4 in Appendix B.)
Let y_w be the sum of these $\dot{x}_w$ numbers *y*. Set $x_w = \dot{x}_w + y_w$.

(iii) Calculate $\dot{X}_w = \dot{X}_{w-1} + \dot{x}_w$, i.e. the number of responding groups after w waves, and $X_w = X_{w-1} + x_w$, i.e. the number of individual respondents after w waves.

(iv) If $w < W$, set $w = w+1$ and repeat from (ii). Otherwise, stop.

Let $\overline{\dot{X}_W}$ resp. $\overline{X_W}$ be the average of the S randomly generated values $\dot{X}_W$ resp. X_W. These are estimates of the theoretical means $E(\dot{X}_W)$ resp. $E(X_W)$.

This global simulation has been implemented in SPSS syntaxes. As an example, we now apply the procedure for the following choice of parameters: m = 3000, W = 4, C = 2, p = 12%. Then G = 8 and n = 24000. S = 1000 simulations were executed. The complete output produced by our SPSS program is as follows:

```
Run MATRIX procedure:

Number of groups (m)
  3000
Number of waves (W)
  4
Number of contacts per wave and per group (C)
  2
Size of groups (G=W*C)
  8
Initial sample size (n=m*W*C)
  24000
Respons probability (%) (p)
  12
Number of simulations (S)
  1000

------ END MATRIX -----

SEED = 416.459.956
```

Descriptives for CRESP_I Total number of respondents

		Statistic	Std. Error
Mean		2042,36	,966
95% Confidence Interval for Mean	Lower Bound	2040,47	
	Upper Bound	2044,26	
5% Trimmed Mean		2042,44	
Median		2043,00	
Variance		932,216	
Std. Deviation		30,532	
Minimum		1944	
Maximum		2125	
Range		181	
Interquartile Range		43,00	
Skewness		-,030	,077
Kurtosis		-,250	,155

Descriptives for CRESP_C Total number of responding groups

		Statistic	Std. Error
Mean		1920,16	,843
95% Confidence Interval for Mean	Lower Bound	1918,51	
	Upper Bound	1921,82	
5% Trimmed Mean		1920,18	

Median		1921,00	
Variance		710,644	
Std. Deviation		26,658	
Minimum		1831	
Maximum		1994	
Range		163	
Interquartile Range		37,00	
Skewness		-,032	,077
Kurtosis		-,288	,155

Thus we see that, for parameter values as given here before, an estimate $\overline{X_W}$ for the expected number $E(X_W)$ of individual responses is 2042.36. With $S = 1000$ replications, the standard error of this estimate is very small (0.966), leading to a very narrow confidence interval for the estimate $\overline{X_W}$. The estimate for the variance $V(X_W)$ equals 932.216 and the corresponding estimate of the standard deviation $\sqrt{V(X_W)}$ equals 30.53. This relatively small standard deviation implies that the expected number of respondents $E(X_W)$ can be predicted very accurately.

Similar remarks can be made about the distribution of the number $\dot{X}_W$ of responding groups, with estimate 1920.16 for its mean $E(\dot{X}_W)$.

The same results are obtained through exact calculations based on the formulas presented in Property C.2 in Appendix C. The numerical results of these exact calculations are shown in the table in Appendix E, in column "Simulation 3" for the above set of parameters, where also the individual and global simulations obtained with the SIR procedure are shown.

Why then that global simulation procedure? We developed this procedure at the moment that exact formulas were not yet available. Now we can, first of all, use simulations to verify numerically the mathematical model and exact formulas. Secondly, the above table gives more information about the distributions of X_W and $\dot{X}_W$, through estimates of parameters such as minimum, maximum, skewness, kurtosis, … This provides interesting information as long as we don't have mathematical formulas for these parameters and/or a full mathematical description of the distributions of X_W and $\dot{X}_W$.

4 Effects of parameters on the response sample; the HPS case

A number of exact calculations and simulations have been performed in order to find an appropriate set of parameters that would finally define the sampling plan for HPS. After having recorded the real situation for HPS, we again verified the procedures taking the HPS reality into account. The results of these simulations and exact calculations are shown in the table in Appendix E. The main purpose of the subsequent discussion is to show that we indeed have met the main objectives put forward in section 1.2.

4.1 *Towards a SIR procedure for HPS*

First of all, the parameter R (number of desired respondents) is fixed from now on: R = 1400 (see section 1.2).

Obviously, time is a serious constraint in the context of the SIR procedure. IPH stated that the study should take at most 4 months. Since in each wave invited persons should have enough time to answer, since IPH should have enough time to record response and transfer this information to Statistics Belgium, and since Statistics Belgium should have enough time to prepare the mailing for a possible next wave, the number of planned waves W should be kept rather small. Although time aspects have not been modelled explicitly, they have implicitly been taken into account when proposing different sets of parameters, as it will become clear hereafter.

IPH initially proposed the following set of parameters: (m, W, C, p) = (1400, 9, 1, 11%). In fact, the value 9 for W was implied by the values for C and p: $9 \cong 1/p = 1/0.11$, and if 9 contacts (C=1 per wave) are made per group, about 1 respondent on average is found in each group, so that R would be reached. However, our simulations and calculations (see column "Simulation 1" in the table of Appendix E) clearly show that only about 909 respondents are expected, which is far below R = 1400. This is because in each group not 1 respondent on average, but *at most* 1 (C=1) respondent is obtained (see 'replacement procedure' described in section 3.2)[v]. Actually, theory clearly shows that an infinite number of waves would have to be organised to reach this R. Impossible! The problem with the proposed set of parameters is that $m = R$ and C = 1: then *at most* R respondents can be found, whatever the value of W and p. Another problem with this parameter set is that W=9 waves take really too much time, taking into account the constraint that the survey should not take more than 4 months after the initial contacting and the fact that 1 wave takes at least 3 weeks.

Therefore, the following alternative set of parameters has been proposed: (m, W, C, p) = (3000, 5, 1, 12%). The rate of 12% was argued by the fact that Statistics Belgium generally experiences a response of 12% for its (voluntary) HBS. Of course, this value was believed to be an underestimate for the real response rate for HPS, e.g. because of the much lower burden for respondents, but all parties accepted a safe underestimate (IPH initially suggested working with a response rate of 10%). Our simulations and calculations (see column "Simulation 2" in the table of Appendix E) show that the new proposal is not too bad: R can be reached after 5 waves. In reality R might be reached earlier, if indeed the true p is considerably larger than 12%.

Nevertheless, after the first theoretical and simulation results had been obtained by Statistics Belgium, it was realised that C could be 2 or more.

Thus, IPH proposed a compromise: increase the number C of contacts per group and wave from 1 to 2, but decrease the number of waves. Hence, a final set of parameters

[v] Unless late positive response is obtained (i.e. from one of the previous waves, while a new wave is already running) and still recorded for groups that were not yet inactivated.

was then obtained quite fast: (m, W, C, p) = (3000, 4, 2, 12%), as can be seen in column "Simulation 3" of the Appendix E table.The rather large value for m was justified by the required representativity of the (final) respondent sample. Indeed, this m guarantees that a lot of different PSUs are selected into the sample, and since PSUs are made up of a quite homogeneous set of individuals (with respect to domicile, age and sex), the initial sample of individuals as well as the final respondent sample (given a uniform p) could also be expected to be representative of the population. This statement is actually justified by the percentage distributions shown in the table in Appendix E: the distributions based on the initial sample and on the respondent sample, generated by SIR simulations for given parameter sets, are all close to the population distributions.

Thus, a final set of parameters was nearly found. Given the results of various simulations, it was clear that only 4 instead of 5 waves had to be planned. Even 3 waves would be enough to reach R, but a fourth wave was retained for safety reasons. Moreover, there was no need to reduce the number W of waves further, since the organisation of 4 waves, each of them taking 4 or 5 weeks, still fitted into the available total period of about 4 months for the whole study.

4.2 *Applying the SIR procedure to HPS*

Finally, planning of the SIR procedure for HPS has then been based on the parameter set (m, W, C) = (3000, 4, 2) ("Simulation 3" in the table of Appendix E). After drawing the initial sample of n = 24000 individuals, mC = 6000 selected individuals were assigned to wave 1, and another 6000 individuals to each of the remaining 3 waves. Statistics Belgium then prepared and sent the mailing of 6000 advance letters (invitations to participate) that make up the 1st wave. The 6000 potential respondents were invited to answer positively to IPH directly: they thus implicitly informed IPH that it could use their addresses for sending the test set. Privacy regulations have that way been satisfied.

Among the 6000 individuals, some 34 % (2036) persons answered positively: they were willing to participate in the HPS.

Representativity of this resulting respondent sample was examined (figures are not shown), and the conclusion was that the objectives had been met.

IPH sent test sets to all these respondents. At present, 1835 test sets have been returned to IPH, 1830 of which were usable. This means that a dropout of 206/2036 = 10.12% has been observed, or that the final overall response rate is 1830/6000 = 30.5%.

The marginal distributions of age, sex and province (stratum) in the final respondent sample are shown in the last column in the table in Appendix E. Apart from age classes 00-14 and 65+, acceptably small differences between observed and population percentages occur. IPH is satisfied with this result. Notice further that, as already pointed out, corrections for non-response and calibration to population information are always suitable techniques for adjusting the initial weights before estimation of

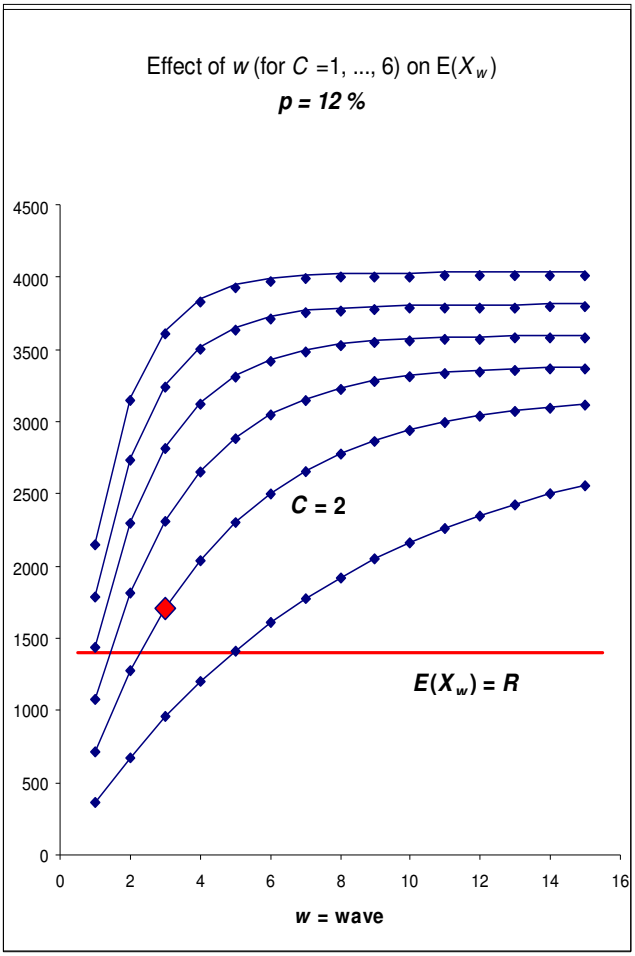
population parameters and analysis of the data starts. These issues are however not considered further in this text.

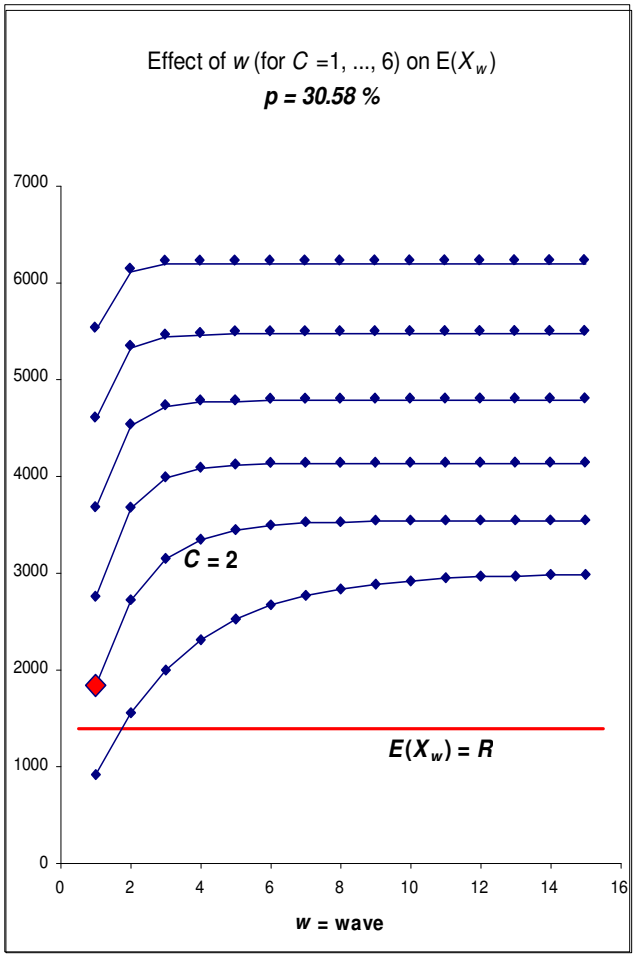
Being satisfied with the results after wave 1, and stopping rule 1 being satisfied, IPH decided to stop the procedure after wave 1.

4.3 *Post SIR simulations for HPS*

Our experiments can be closed by a few more simulations and exact calculations for HPS, based on real life (response) experience recorded in wave 1.

The following 2 plots show the influence of parameters C and w on the expected cumulative number of respondents $E(X_w)$ after w waves, as calculated by the 1st formula in Property C.2 (Appendix C). The 1st plot holds for response rate p=12% (*assumed* response rate used when planning the survey), the second one for p=30.58% (*observed* response rate after 1 wave, after dropout, but including unusable returned test sets). Both plots suppose m=3000 as number of groups in the sample and have R=1400 as the desired number of respondents (indicated by the red horizontal line). The latter must be interpreted as the number of respondents that must be reached before stopping the contacting procedure. The 6 curves (seen from bottom to top) within each plot correspond to values 1, … 6 of the parameter C, the number of contacts per active group and per wave.





Such plots allow accurate planning and evaluation of the sampling design *and* response modelling. The 1st plot was used for the *planning* of the procedure of sampling, inviting and response recording and shows that for C=1 certainly 5 waves are needed to reach the threshold of 1400 respondents, whereas with C=2, only 3 waves are needed: see the point marked by the large red diamond. For safety reasons, we finally decided to plan W=4 waves. In the 2nd plot, the curve for C=2 shows that already sufficient respondents will be obtained after 1 wave (see red diamond point), *which perfectly corresponds to what was observed in practice*. With more precise knowledge of the response rate p, we could even have chosen C=1 and noticed that 2 waves would then have been sufficient for reaching the threshold R. In that case, the number of observed responses would have been even closer to the desired number R=1400.

Notice that completely similar curves would result for other values of m, as that value is merely a proportionality constant, which does not change the shapes of the curves.

The results from the above plots correspond to those summarized in the table of Appendix E (containing the results from both individual and global simulations). "Simulation 4" shows the result for parameters (m, W, C, p) = (3000, 4, 2, 34%) and "Simulation 5" (p corresponding nearly to p=30.58 in the 2nd of the above plots) shows the results for (m, W, C, p) = (3000, 4, 2, 31%). This corresponds to using the same values for parameters m, W and C as for the actual planning of HPS ("Simulation 3", corresponding to the 1st of the above plots) but using response rates p close to *observed* rates (instead of the 12% from the planning phase), one being the initial response rate before test sets are sent and the other one being the response rate after test sets have been returned (real participation after dropout). The figures under "Simulation 4" and "Simulation 5" in the table show that with a better a priori guess about the true response rate, one might indeed have reduced the planned number of waves to 1 (or 2 for safety reasons) and that a similar number of final respondents would have been obtained.

It is seen from Appendix E (see the last 4 columns in the table) that observed experience for HPS is remarkably close to calculations based on exact mathematical formulas. Simulation 4 shows that with (m, C, p) = (3000, 2, 34%), the observed number of respondents after 1 wave (*before* test sets are sent) is 2036 (observed response rate p=33.93), while the expected number is 2040.0, with a standard deviation of 36.69. Simulation 5 shows that with (m, C, p) = (3000, 2, 31%), the observed number of respondents after 1 wave (*after* test sets are sent, thus taking dropout into account) is 1830 (observed response rate p=30,5), while the expected number is 1860.0, with a standard deviation of 35.82.
With (m, C, W, p) = (3000, 2, 1, 30.5%), where 30.5% = 1830/6000, the expected number of respondents after 1 wave is exactly 1830 (just as it was observed), with a standard deviation of 35.66. Global simulation (section 3.5), with S = 1000, gives an estimated mean of 1833.33 with standard deviation 35.69 after 1 wave, one individual simulation (section 3.4) gives 1825 respondents after 1 wave. A similar correspondence between expected number of respondents (and its standard deviation),

observed number of respondents, global and individual simulations occur for (m, C, W, p) = (3000, 2, 1, 33.93%), where 30.93% = 2036/6000.

Thus, theory and observed reality perfectly match!

It is possible to change, based on real life experience, other parameters of the SIR procedure, but that will definitely not change the present observed reality. Next time a HPS (or similar survey) is being organised, this experience can surely be taken into account, and planning of the survey could be considerably refined!

5 Future work

Our research has clearly revealed that the SIR procedure might be a valuable alternative for dealing with non-response in the planning phase of a survey. One topic for future work is definitely a thorough comparison of SIR with the other methods presented in section 2 and currently used at Statistics Belgium for several of its sample surveys. It might lead to revision of our HBS, LFS and other household or individual surveys, as well as business surveys. Of course, several practical issues, such as interviewer workload, cost, and – not the least important – timing, require considerably more investigation. But an appropriate management tool, to be used when planning a survey and when data are collected, is indispensable nowadays, and we believe that a tool based on the SIR procedure could be a suitable alternative.

It has been shown that a sound mathematical framework and a flexible simulation tool are very helpful both in understanding the mechanics of the SIR procedure and in detailed planning of a survey. Some refinements of the model deserve attention: it should be possible to work with varying response rates (varying over strata, PSUs…), it would be nice to be able to use observed response in planning for future waves, etc.

However, although response rates are certainly varying across PSUs, simply since these are different to some extent with respect to some background characteristics (i.e. variables defining strata and PSUs), observation matches very well with simulations and exact calculations based on a simplifying model with uniform response rate p. This questions the need for more refinements in the mathematical model, as far as only the total numbers $E\left(X_W\right)$ and $E\left(\dot{X}_W\right)$ are of interest. Of course, if these parameters need to be known in advance at stratum level, or at PSU level, an extension (be it quite straightforward) of the present model is inevitable.

Finally, some sampling design supporting features should be incorporated into general software for sampling. SIR, which encompasses traditional approaches for special choices of its parameters, might be considered in that respect. The fact that fast global simulations, or exact calculations when formulas become available, are possible, is worth being considered!

Bibliography

Beutels, M. *et al* (1997) Prevalence of Hepatitis A, B and C in the Flemish Population, *European Journal of Epidemiology*, Vol. 13, pp.215-280.

Dawagne, J.-M. (2004) *L'Estimation de la Variance dans les Enquêtes par Sondage et Application à l'Enquête sur les Forces de Travail*, Mémoire de stage présenté en vue de la nomination au grade d'Actuaire. Brussels: Statistics Belgium.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New-York: Springer.

Slock, P. (2004) *Sampling Methods and Techniques at Statistics Belgium*, Report submitted for promotion to the degree of Actuary. Brussels: Statistics Belgium.

Vanderhoeft, C. (2004) *Notes on the Mathematics of an Incremental 2-Stage Sampling Design*, Manuscript. Brussels: Statistics Belgium.

Biography

Peter Slock obtained the degrees of "Civil Engineer in Applied Mathematics and Computer Sciences" (KULeuven 1995) and "Master of Statistics" (KULeuven 2000). Since 2001 he's working as methodologist and quality manager at Statistics Belgium (*Statbel*) in the bureau of methodology and co-ordination. He mainly gained experience in applying statistical methods to various statistical domains. Last year, he worked more profoundly on topics of survey sampling, focusing on various sampling methods and techniques and their application to surveys at Statistics Belgium, which is becoming his main area of interest and activity. He's also interested in index theory and seasonal adjustment of time series of indices, as well as in metadata and other quality related issues of statistics.

Address: *Peter.Slock@statbel.mineco.fgov.be*, Statistics Belgium, 44 Leuvenseweg, B-1000 Brussels, Belgium

Camille Vanderhoeft, Master in Actuarial Sciences (VUBrussels 1980) and PhD in Mathematics (VUBrussels 1989). Working at Statistics Belgium since 1997. Main interest is in mathematical aspects of survey methodology, and in development of general software for various phases of the survey process. Still affiliated to the VUBrussels, where teaching a basic course on statistics in a management programme.

Address: *Camille.Vanderhoeft@statbel.mineco.fgov.be*, Statistics Belgium, 44 Leuvenseweg, B-1000 Brussels, Belgium

Sophie Quoilin, Medical Doctor (1993) in Public Health. Currently working at the Epidemiology Unit of the Scientific Institute of Public Health in Brussels, Belgium. In charge of different infectious disease projects such as hepatitis, Creutzfeldt-Jakob disease, zoonotic diseases. Also interested in tropical diseases and disability issues.

Address: *Sophie.Quoilin@iph.fgov.be*, Scientific Institute of Public Health, 14 J. Wytsmanstraat, B-1050 Brussels, Belgium

Appendix A The binomial and related distributions

The results in this and the following two appendices are useful for justifying the simulation procedures proposed in the main text.

Working hypotheses

1. All individuals have the same response probability *p*.
2. The response probability *p* doesn't change over waves.
3. Response is independent between individuals, within and between groups, and between PSUs.

Notation
Let $x \sim B(C, p)$ denote a binomially distributed random variable with parameters *C* (the number of trials in a binomial experiment) and *p* (the probability of success in each trial). Let $f(k;C,p) = \binom{C}{k} p^k (1-p)^{C-k}$. Notice that $E(x) = Cp$ and $V(x) = Cp(1-p)$.

Property A.1

$$\sum_{k=1}^{C} k \cdot f(k;C,p) = Cp$$

$$\sum_{k=1}^{C} k^2 \cdot f(k;C,p) = Cp(1-p) + C^2 p^2$$

Proof
Follows easily from the above-mentioned properties of the binomial $B(C, p)$ distribution.

■

Definition A.1 (*Group response and additional individual response*)
From $x \sim B(C, p)$, derive the following random variables:

$$\dot{x} \begin{cases} = 0 & \text{iff } x = 0 \\ = 1 & \text{iff } x > 0 \end{cases} \quad \text{and} \quad y \begin{cases} = 0 & \text{iff } x \le 1 \\ = x - 1 & \text{iff } x \ge 2 \end{cases}$$

If *x* is the number of respondents in a group of *C* individuals, each having response probability *p*, then $\dot{x}$ can be interpreted as the response indicator for the group, and *y* as the *additional (individual) response*. Notice that $y = x - \dot{x}$.

For $C = 1$, we get $\dot{x} \equiv x$ and $y \equiv 0$. This special case will not be mentioned anymore explicitly hereafter, since all results are valid for $C = 1$ too.

Property A.2 (*Distribution of group response*)

$\dot{x} \sim B(1, \dot{p})$ with $\dot{p} = 1-(1-p)^C$, whence $E(\dot{x}) = \dot{p}$ and $V(\dot{x}) = \dot{p}(1-\dot{p})$.

Proof Straightforward. ■

Property A.3 (*Distribution of additional response*)

The distribution of the additional individual response *y* is as follows:

$$P(y=0) = f(0;C,p) + f(1;C,p)$$

and $$P(y=k) = f(k+1;C,p) \quad \text{for} \quad k = 1,...,C-1.$$

Mean and variance of *Y* are:

$$E(y) = Cp - \dot{p} = Cp - 1 + (1-p)^C$$

and $V(Y) = Cpq + \dot{p}\dot{q} - 2Cp\dot{q}$, with $q = 1-p$ and $\dot{q} = 1-\dot{p}$.

Proof

By Definition A.1:

$$P(y=0) = P(x=0 \vee x=1) = f(0;C,p) + f(1;C,p)$$

and $$P(y=k) = P(x=k+1 \wedge \dot{x}=1) = P(x=k+1) = f(k+1;C,p) \text{ for } k>0.$$

The mean of *Y* is easily found:

$$E(y) = E(x-\dot{x}) = E(x) - E(\dot{x}) = Cp - \dot{p}.$$

The covariance of *x* and $\dot{x}$ is:

$$\begin{aligned} Cov(x,\dot{x}) &= E(x\dot{x}) - E(x)E(\dot{x}) \\ &= \sum_{k=0}^{C} k \cdot 0 \cdot P(x=k \wedge \dot{x}=0) + \sum_{k=0}^{C} k \cdot 1 \cdot P(x=k \wedge \dot{x}=1) - E(x)E(\dot{x}) \\ &= \sum_{k=1}^{C} k \cdot P(x=k \wedge \dot{x}=1) - E(x)E(\dot{x}) = \sum_{k=1}^{C} k \cdot P(x=k) - E(x)E(\dot{x}) \\ &= E(x) - E(x)E(\dot{x}) = Cp(1-\dot{p}) = Cp\dot{q} \end{aligned}$$

The variance of *y* then follows easily:

$$\begin{aligned} V(y) &= V(x-\dot{x}) = V(x) + V(\dot{x}) - 2Cov(x,\dot{x}) \\ &= Cpq + \dot{p}\dot{q} - 2Cp\dot{q} \end{aligned}$$

■

Property A.4 (*Conditional distribution of additional response*)

The conditional distribution of *y* given group response $\dot{x}$ is specified as follows:

$$P(y=0 \,|\, \dot{x}=0) = 1$$

and $$P(y=k \,|\, \dot{x}=1) = f(k+1;C,p)/\dot{p} \quad \text{for} \quad k = 0,...,C-1.$$

Conditional means and variances of *y* are:

$$E(y \,|\, \dot{x}=0) = 0 \quad \text{and} \quad V(y \,|\, \dot{x}=0) = 0$$

and $$E(y \,|\, \dot{x}=1) = \dot{p}^{-1}(Cp - \dot{p}) \quad \text{and} \quad V(y \,|\, \dot{x}=1) = Cp\dot{p}^{-1}(q - Cp\dot{q}\dot{p}^{-1})$$

Proof

The distribution of *y* given $\dot{x} = 0$, and its mean and variance are obvious.

For *k* = 0, the conditional probability $P(y=k \,|\, \dot{x}=1)$ equals:

$$P\left(y=0\middle|\dot{x}=1\right)=\frac{P\left(y=0\wedge\dot{x}=1\right)}{P\left(\dot{x}=1\right)}=\frac{P\left(x=1\right)}{P\left(\dot{x}=1\right)}=\frac{f\left(1;C,p\right)}{\dot{p}}.$$

For $k=1,...,C-1$, we get, using Property A.3:

$$P\left(y=k\middle|\dot{x}=1\right)=\frac{P\left(y=k\wedge\dot{x}=1\right)}{P\left(\dot{x}=1\right)}=\frac{P\left(y=k\right)}{P\left(\dot{x}=1\right)}=\frac{f\left(k+1;C,p\right)}{\dot{p}}.$$

Notice that $E\left(y^{\alpha}\right)=E\left(y^{\alpha}\middle|\dot{x}=0\right)P\left(\dot{x}=0\right)+E\left(y^{\alpha}\middle|\dot{x}=1\right)P\left(\dot{x}=1\right)=E\left(y^{\alpha}\middle|\dot{x}=1\right)\dot{p}$, for $\alpha = 1, 2, \ldots$, whence, with Property A.3:

$$E\left(y\middle|\dot{x}=1\right)=\dot{p}^{-1}E\left(y\right)=\dot{p}^{-1}\left(Cp-\dot{p}\right)=Cp\dot{p}^{-1}-1.$$

Similarly: $E\left(y^{2}\middle|\dot{x}=1\right)=\dot{p}^{-1}E\left(y^{2}\right)$, whence:

$$\begin{aligned}V\left(y\middle|\dot{x}=1\right)&=E\left(y^{2}\middle|\dot{x}=1\right)-E^{2}\left(y\middle|\dot{x}=1\right)=\frac{E\left(y^{2}\right)}{\dot{p}}-\frac{E^{2}\left(y\right)}{\dot{p}^{2}}\\&=\frac{E\left(y^{2}\right)}{\dot{p}}-\frac{E^{2}\left(y\right)}{\dot{p}}+\frac{E^{2}\left(y\right)}{\dot{p}}-\frac{E^{2}\left(y\right)}{\dot{p}^{2}}=\frac{V\left(y\right)}{\dot{p}}-\frac{E^{2}\left(y\right)\left(1-\dot{p}\right)}{\dot{p}^{2}}\\&=\frac{Cpq+\dot{p}\dot{q}-2Cp\dot{q}}{\dot{p}}-\frac{\left(Cp-\dot{p}\right)^{2}\dot{q}}{\dot{p}^{2}}\\&=Cpq\dot{p}^{-1}+\dot{q}-2Cp\dot{q}\dot{p}^{-1}-\frac{\dot{q}}{\dot{p}^{2}}\left(C^{2}p^{2}-2Cp\dot{p}+\dot{p}^{2}\right)\\&=Cpq\dot{p}^{-1}-C^{2}p^{2}\dot{q}\dot{p}^{-2}=Cp\dot{p}^{-1}\left(q-Cp\dot{q}\dot{p}^{-1}\right)\end{aligned}$$

■

Appendix B Simulating response

In this appendix, methods for generating random numbers from binomial and related distributions are presented. Some of these have been used in the global simulation procedure, as discussed in section 3.5.

SPSS is used for implementing simulation of (non-)response. The package has built-in functions for generating random numbers from standard distributions, such as the binomial. For non-standard distributions, random numbers are obtained indirectly from standard distributions. Random number generation (*RNG*) methods B.1 to B.4 are useful for simulation of response within groups of size C.

RNG B.1 *Simulation of individual response x*

A random number x from the $B(C, p)$ distribution can be generated as follows:

Generate $x = \text{rv.binom}(C, p)$.

RNG B.2 *Simulation of group response* $\dot{x}$

A random number $\dot{x}$ from the $B(1, \dot{p})$ distribution, with $\dot{p} = 1-(1-p)^C$, can be generated in two ways:

Method 1 : Generate $\dot{x} = \text{rv.binom}(1, \dot{p})$.
Method 2 : Generate $x = \text{rv.binom}(C, p)$; compute $\dot{x} = \min(x,1)$.

RNG B.3 *Simulation of unconditional additional response y*

A random number y from the unconditional distribution of additional response, as in Property A.3, can be generated in two ways:

Method 1 : Generate $x = \text{rv.binom}(C, p)$; compute $\dot{x} = \min(x,1)$ and $y = x - \dot{x}$.
Method 2 : Generate $x = \text{rv.binom}(C, p)$; compute $y = \max(0, x\text{-}1)$.

RNG B.4 *Simulation of conditional additional response y, given group response* $\dot{x} = 1$

A random number y from the conditional distribution of additional response given $\dot{x} = 1$, as in Property A.4, can be generated as follows:

Generate $x = \text{rv.binom}(C, p)$ until $x > 0$; then compute $y = x - 1$.

Appendix C Cumulative individual response

Definition C.1
Let X_w^{ij} be the cumulative number of respondents after wave *w* in group *ij* (irrespective of the wave – before, at or after wave *w* – in which group *ij* responds, if it ever responds at all).
Let $X_w = \sum_{ij} X_w^{ij}$, where the summation is over all *m* (selected) groups. X_w is the *total cumulative individual response* after wave *w*.

Property C.1
Assume that the SIR procedure continues until stopping rule 2 is satisfied; let $w \le W$.
The distribution of X_w^{ij} is defined as:

$$P\left(X_w^{ij}=0\right)=(1-p)^{wC}=(1-\dot{p})^w$$

and $$P\left(X_w^{ij}=k\right)=\frac{1-(1-\dot{p})^w}{\dot{p}} f\left(k;C,p\right) \text{ for } k=1,...,C.$$

Its mean and variance are resp.

$$E\left(X_w^{ij}\right)=\frac{1-(1-\dot{p})^w}{\dot{p}}Cp$$

and $$V\left(X_w^{ij}\right)=\frac{1-(1-\dot{p})^w}{\dot{p}}Cp(1-p)+\frac{1-(1-\dot{p})^w}{\dot{p}}\left(1-\frac{1-(1-\dot{p})^w}{\dot{p}}\right)C^2p^2.$$

Proof
The probability that among *C* individuals no one responds equals $(1-p)^C=1-\dot{p}$.
The expression for $P\left(X_w^{ij}=0\right)$ follows immediately.
Finding *k* respondents in wave *v* means that in waves 1 to $v-1$ no respondent is found, while *k* respondents are found among *C* individuals in wave *v*. The probability of this event is $(1-\dot{p})^{v-1} f\left(k;C,p\right)$. Hence:

$$P\left(X_w^{ij}=k\right)=\sum_{v=1}^{w}(1-\dot{p})^{v-1} f\left(k;C,p\right)=\frac{1-(1-\dot{p})^w}{\dot{p}} f\left(k;C,p\right).$$

The mean is easily calculated:

$$E\left(X_w^{ij}\right)=\sum_{k=0}^{C}k\cdot P\left(X_w^{ij}=k\right)=\sum_{k=1}^{C}k\cdot P\left(X_w^{ij}=k\right)$$
$$=\frac{1-(1-\dot{p})^w}{\dot{p}}\sum_{k=1}^{C}k\cdot f\left(k;C,p\right)=\frac{1-(1-\dot{p})^w}{\dot{p}}Cp,$$

where the last equality follows from Property A.1. The second central moment also follows easily from Property A.1:

$$E\left(\left(X_w^{ij}\right)^2\right)=\sum_{k=0}^{C}k^2\cdot P\left(X_w^{ij}=k\right)=\sum_{k=1}^{C}k^2\cdot P\left(X_w^{ij}=k\right)$$
$$=\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}\sum_{k=1}^{C}k^2\cdot f\left(k;C,p\right)$$
$$=\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}\left(Cp\left(1-p\right)+C^2p^2\right)$$

The variance thus equals:

$$V\left(X_w^{ij}\right)=E\left(\left(X_w^{ij}\right)^2\right)-E^2\left(X_w^{ij}\right)$$
$$=\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}\left(Cp\left(1-p\right)+C^2p^2\right)-\left(\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}Cp\right)^2$$
$$=\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}Cp\left(1-p\right)+\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}\left(1-\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}\right)C^2p^2$$

■

Because of independence between groups, the distribution of X_w is the distribution of the sum of *m* independent and identically distributed variables X_w^{ij}, whose distributions are as in Property C.1. The full distribution is hard to calculate, but its main parameters are easily derived; the result is stated in the following Property C.2.

Property C.2
Assume that the SIR procedure continues until stopping rule 2 is satisfied; let $w\le W$. The mean and variance of X_w are resp.

$$E\left(X_w\right)=m\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}Cp$$

and $$V\left(X_w\right)=m\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}Cp\left(1-p\right)+m\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}\left(1-\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}\right)C^2p^2.$$

■

Remember that $\dot{p}=1-\left(1-p\right)^C$, whence $\frac{1-\left(1-\dot{p}\right)^w}{\dot{p}}=\frac{1-\left(1-p\right)^{wC}}{1-\left(1-p\right)^C}=\frac{1-q^{wC}}{1-q^C}$. This leads to alternative expressions for the mean and variance in Property C.2:

$$E\left(X_w\right)=mCp\frac{1-q^{wC}}{1-q^C},$$
$$V\left(X_w\right)=mCp\frac{1-q^{wC}}{1-q^C}\left[q+Cp-Cp\frac{1-q^{wC}}{1-q^C}\right].$$

For $w = 1$, $\frac{1-q^{wC}}{1-q^{C}} = 1$ and these expressions simplify as follows:

$$E\left(X_1\right) = mCp = E\left(B\left(mC, p\right)\right),$$

$$V\left(X_w\right) = mCpq = V\left(B\left(mC, p\right)\right).$$

Finally, notice that for $w > 1$, the "correction factor" $\frac{1-q^{wC}}{1-q^{C}} \geq 1$.

Appendix D Hypothetical example illustrating initial sampling, inviting to participate and inactivating groups

Example 4 selected PSUs i ($m' = 4$); W: at least 4 waves; assume that the procedure stops after wave $w = 4$

						Contacted SSUs							
PSU	Multiplicity	Group	Selected SSUs per group	Selected SSUs per wave and group	Wave in which the group finally responds	Wave	Wave	Wave	Wave	Wave in which the group responds, or w	Number of SSUs finally contacted	Number of SSUs finally contacted	Initial weight
i	m_i	j	G	$C = G/W$	w_{ij}	1	2	3	4	$\tilde{w}_{ij}$	$\tilde{w}_{ij}C$		$d_{ijk}^{(w)}$
1	1	1	G	C	1	C	0	0	0	1	C	C	a/C
2	1	1	G	C	2	C	C	0	0	2	$2C$	$2C$	$a/(2C)$
3	2	1	G	C	2	C	C	0	0	2	$2C$	$3C$	$a/(2C)$
		2	G	C	1	C	0	0	0	1	C		a/C
4	4	1	G	C	1	C	0	0	0	1	C	$12C$	a/C
		2	G	C	3	C	C	C	0	3	$3C$		$a/(3C)$
		3	G	C	4	C	C	C	C	4	$4C$		$a/(4C)$
		4	G	C	∞	C	C	C	C	4	$4C$		$a/(4C)$
Total	$m = 8$		$n = 8G$	$n/W = 8C$							$18C$	$18C$	

$\mathcal{G}^{(1)} = \{11, 21, 31, 32, 41, 42, 43, 44\}$ with size $m = 8$

$\mathcal{G}^{(2)} = \{21, 31, 42, 43, 44\}$ with size $m - \dot{X}_1 = 5$

$\mathcal{G}^{(3)} = \{42, 43, 44\}$ with size $m - \dot{X}_2 = 3$

$\mathcal{G}^{(4)} = \{43, 44\}$ with size $m - \dot{X}_3 = 2$

$\mathcal{G}^{(\infty)} = \mathcal{G}^{(5)} = \{44\}$ with size $m - \dot{X}_4 = 1$

Appendix E: Simulation results, compared with exact calculations and real life case of HPS

		Simulation 1	*Simulation 2*	***Simulation 3***	*Simulation 4*	*Simulation 5*	***Positive responses HPS***	***Real participation HPS***
Parameters	*m*	1400	*3000*	***3000***	*3000*	*3000*	**Initial**	***after dropout***
	W	9	5	**4**	4	4		
	C	1	1	**2**	2	2		
	p	11	12	**12**	34	31	**33.93**	**30.5**
Expected cumulative number of respondents (and standard deviation): exact calculations							***Observed***	
After *W* waves		909.5 (17.85)	1416.8 (27.34)	2043.7 (29.94)	3484.3 (24.94)	3367.9 (25.12)		
Wave *w* after which *R* is reached		∞	5	3	1	1	1	1
After *w* waves		1400.0 (0)	1416.8 (27.34)	1709.3 (30.67)	2040.0 (36.69)	1860.0 (35.82)	***2036***	***1830***
Global simulation (S=1000)		909.2 (18.20)	1416.6 (27.42)	2042.4 (30.53)	3485.2 (25.10)	3369.2 (25.16)	2038.4 (36.39)	1833.3 (35.69)
IND simulation		923	1416	2051	3478	3341	2032	1825
Comparing sampling with population distributions (%) after SIR simulations								
	Population	***Initial*** **Resp.**	***Initial*** **Resp.**	***Initial*** **Resp.**	***Initial*** **Resp.**	***Initial*** **Resp.**	***… before dropout***	***… after dropout***
Age classes								
00-14	**15.8**	*15.7* 15.3	*15.7* 15.0	*15.5* 14.8	*15.2* 15.5	*15.3* 15.2	**19.8**	**19.2**
15-24	**12.0**	*11.9* 12.9	*12.5* 11.7	*12.0* 12.2	*12.6* 12.5	*12.9* 12.4	**11.6**	**11.5**
25-34	**12.9**	*13.2* 13.1	*12.7* 13.6	*12.9* 13.7	*13.1* 12.8	*13.1* 13.2	**11.4**	**10.7**
35-44	**15.9**	*16.5* 17.7	*15.9* 16.4	*15.7* 14.6	*15.2* 14.9	*15.0* 14.9	**16.7**	**16.8**
45-54	**14.4**	*14.8* 14.1	*14.3* 14.8	*15.3* 16.1	*14.7* 14.8	*15.0* 15.1	**17.1**	**17.6**
55-64	**11.4**	*10.7* 9.6	*11.2* 12.1	*11.3* 11.4	*11.9* 11.5	*11.6* 11.6	**12.1**	**12.6**
65+	**17.6**	*17.1* 17.3	*17.8* 16.4	*17.2* 17.2	*17.4* 17.9	*17.1* 17.6	**11.3**	**11.6**
Sex								
Male	**49.6**	*49.1* 49.1	*50.6* 48.4	*49.2* 49.0	*49.5* 49.9	*48.8* 49.7	**47.6**	**48.3**
Female	**50.4**	*50.9* 50.9	*49.4* 51.6	*50.8* 51.0	*50.5* 50.1	*51.2* 50.3	**52.4**	**51.7**
Strata								
Antwerpen	**27.8**	*27.8* 29.1	*27.8* 27.6	*27.8* 27.5	*27.8* 27.8	*27.8* 27.9	**28.3**	**28.7**
West-Vlaanderen	**18.8**	*18.9* 18.4	*18.8* 18.4	*18.8* 18.5	*18.8* 18.5	*18.8* 18.5	**18.8**	**18.4**
Oost-Vlaanderen	**22.8**	*22.8* 21.8	*22.8* 23.0	*22.8* 23.3	*22.8* 23.2	*22.8* 22.9	**22.3**	**22.0**
Limburg	**13.4**	*13.4* 14.0	*13.4* 13.8	*13.4* 13.4	*13.4* 13.2	*13.4* 13.7	**15.3**	**15.4**
Vlaams-Brabant	**17.2**	*17.1* 16.7	*17.2* 17.2	*17.2* 17.3	*17.2* 17.3	*17.2* 17.0	**15.3**	**15.5**
Largest absolute difference		0.7 1.8	1.0 1.3	0.9 1.8	0.7 0.9	0.9 1.0	**6.4**	**6.0**

Informations

La Direction générale Statistique et Information économique relève du SPF Economie, PME, Classes moyennes et Energie. Une de nos missions est de répondre aux besoins des autorités, des entreprises et des citoyens par une information chiffrée sur la situation réelle du pays dans différents domaines d'actualité

Où trouver l'information statistique et économique?

Sur nos sites Internet http://statbel.fgov.be (statistiques) et http://economie.fgov.be (économie)

Dans cinq grandes villes du pays, la Direction générale Statistique et Information économique met à la disposition du public :

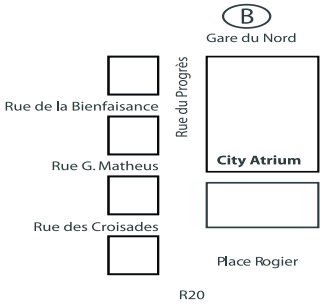
- ◊ Des annuaires et des publications spécialisées ainsi qu'une sélection de disquettes et de cédéroms.
- ◊ Une salle de lecture où il est possible de consulter nos publications, ainsi que celles d'autres ministères ou d'institutions belges et internationales.

Toutes nos bibliothèques sont accessibles les jours ouvrables de 8h30 à 16h30 (Bxl) ou de 9h à 12h et de 13h à 16h (autres).

Bruxelles City Atrium C

Rue du Progrès 50, 1210 Bruxelles
tél. 02/277.55.03 – 02/277.55.04 fax 02/277.55.19
e-mail : info@economie.fgov.be

Train (B) : Gare du Nord
Métro (M) : ligne 2, station Rogier
Trams : 3, 52, 55, 56, 81, 90
arrêts Rogier ou Nord
Bus STIB : 38, 58, 61
arrêts Rogier ou Nord
Bus De Lijn : 318, 351, 358, 410, 526, 554
arrêt Nord



Anvers

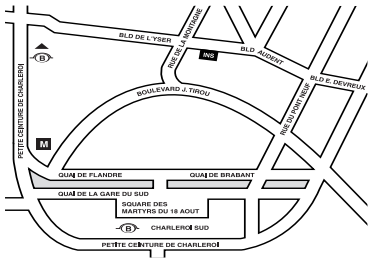
Italiëlei 124 - bus 85, 2000 Antwerpen
tél. 03/229.07.00 fax 03/233.28.30
e-mail : info.antwerpen@economie.fgov.be

Train (B) : Centraal Station
Métro (M) : arrêt Opera
Tram-Bus : accès facile (Fr. Rooseveltplaats)

Charleroi

Tour Biarent, Bd Audent 14/5, 6000 Charleroi
tél. 071/27.44.14 fax 071/27.44.19
e-mail : info.charleroi@economie.fgov.be

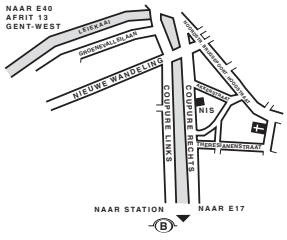
Train (B) : Charleroi Sud, 20 min depuis la gare (Place Buisset, Rue du Collège, Place Charles II, Boulevard Tirou, rue de la Montagne)
Bus : arrêt Tirou
Autoroute : petite ceinture de Charleroi - sortie Gare du Sud
Parking (P) : payant face à l'INS



Gand

Coupure rechts 620, 9000 Gent
tél. 09/267.27.00 fax 09/267.27.29
e-mail : info.gent@economie.fgov.be

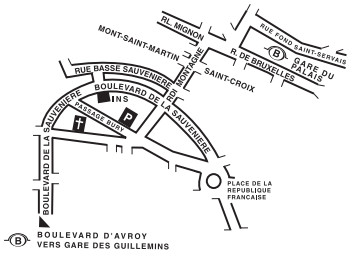
Train (B): Gent St. Pieters
Tram-Bus : 40, 43 arrêt Theresianenstraat
Autoroute: accès aisé par autoroute E40 (sortie N° 13 - Gent - West/Drongen)
Parking (P): au long de la "Coupure Rechts"



Liège

Bd de la Sauvenière 73-75, 4000 Liège
tél. 04/223.84.11 fax 04/222.49.94
e-mail : info.liege@economie.fgov.be

Train (B): Gare des Guillemins ou Gare du Palais
Tram-Bus : (Guillemins) 1 et 4 arrêt Sauvenière
Parking (P): Neujean (à 20 m - même trottoir)
Mercure (en face)



Nous diffusons de nombreux produits qui donnent une image chiffrée de la réalité socio-économique belge. Ces produits, repris dans notre catalogue, sont disponibles auprès de nos centres régionaux ou auprès de notre service de Documentation - vente de Bruxelles. Notre catalogue vous sera envoyé sur simple demande. (voir adresses ci-contre).
Vous trouverez également un extrait de nos données, ainsi que la liste de nos publications sur notre site Internet : http://statbel.fgov.be

Publications générales

Communiqué hebdomadaire

Chaque semaine, nous vous donnons la primeur des dernières statistiques disponibles dans les domaines suivants : Territoire et environnement ; Population ; Société ; Économie et finances ; Agriculture ; Industrie ; Services, commerce et transport.

Chiffres-clés

Cette petite publication explore notre territoire sous ses aspects les plus divers : le climat, l'environnement, la population, la vie sociale, l'économie, les finances, l'agriculture, l'industrie, le transport, la société de l'information… Chiffres-clés 2004 est une brochure gratuite de 50 pages, en couleurs, de format réduit. Vous y trouverez une sélection de la rubrique Statistiques de notre site Internet brossant une vue singulière de l'information statistique disponible en Belgique. Les tableaux sont éclairés par des graphiques et des cartogrammes.

Quelques autres publications

Publications générales

Annuaire de statistiques régionales

Territoire et environnement

Statistique de l'occupation du sol (**disquette**)
Aperçu Environnement - *Annuel*

Population

Mouvement de la population – *Annuel*
Perspectives de population 2000/2050

Société

Enquête sur les budgets des ménages – *Annuel*
Causes de décès - *Annuel*

Économie et finances

Vente de biens immobiliers – *Annuel*
La conjoncture - *Mensuel*

Agriculture

Recensement agricole et horticole
au 15 mai - *Annuel*

Industrie

Production industrielle et construction - *Mensuel*

Commerce, services et transports

Statistiques mensuelles du transport – *Mensuel*
Commerce intérieur - *Annuel*

Achevé d'imprimer
par l'imprimerie de la
Direction générale Statistique
et Information économique
B-1000 Bruxelles

Janvier 2007