2023 Joint Statistical Meetings of the American Statistical Association
Survey Research Methods Section, Metro Toronto Convention Center, Ontario, Canada, August 5-10

Cost Reduction for Big Data Exploration and Pertinent Knowledge Extraction (Part 1)

A. Demnati

Independent Researcher, Ottawa, Canada, Abdellatif_Demnati@msn.com

Abstract

It is particularly difficult to analyze digital observations due to the enormous amount of data per any given time period. The volume of data initially presents with the problem of interception followed by the issue of storage and analysis which all needs to fall within an appropriate infrastructure cost. The lack of pertinent information, extracted from big data, deprives both the evaluation of the use and the performance of the network and ultimately, the possibility of offering an effective service. This study looks to explore intercepted big data and how one may extract pertinent knowledge while minimizing the process cost. The process of analyzing big data is seen as an adaptive two-phase sampling design; the first-phase sample can be both stored within the available infrastructure and used as a sampling frame for each type of potential knowledge that can be extracted. Calibration of multiple sets of weights is completed to ensure the consistency of the estimates with constant quantities as well as between estimates using different or same sets of weights. We examine this proposed approach and the results of an illustrative example are presented.

Key Words: Adaptive two-phase sampling, Digital network, Job shop scheduling, Multiweights calibration, Poisson sampling, Sample size determination.

1. Introduction

The development of the Internet and the increased availability of digital data has elicited a need for data exploration and the extraction of pertinent knowledge through estimation. Digital data (known as big data) is complex and costly to intercept, store, and analyse. For example, suppose you have to store and process one petabyte of data per period of time such as a day or an hour, where a petabyte is E15 bytes of digital information and a byte is eight binary digits long. To depict big data, it is common to use the "5Vs": (1) Velocity: the data is generated quickly in real time; (2) Volume: the amount of data per period of time; which complicates its manipulation using available infrastructure; (3) Variety: data sources and types are diverse; (4) Value: the potential knowledge that can be extracted from the data; and (5) Veracity: the level of inherent quality in big data and the likelihood of being error-prone.

Analyses of a network usage and performance first require the interception of data transmitted between network computers and between others' networks, possibly hidden from users' population. Once intercepted, given that any relevant information present in the data is unknown to the analysts, the challenge consists therefore of (1) gaining insights into the intercepted data at low cost, and (2) defining the type of pertinent information that can be extracted

(i.e., defining the parameter of interest $\boldsymbol{\theta}$ among a set $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ of K potential parameters that can be estimated given the data at hand). Traditional sampling designs can be used under available computing infrastructure. However, there is no guideline on how to select the sample and how to choose the hidden though available, pertinent parameter. Unlike big data, in survey data situations, the parameter of interest $\boldsymbol{\theta}$ is well defined prior to the questionnaire design and data collection and the sample size is derived with respect to the parameter of interest.

In an attempt to analyse intercepted big data while minimizing costs, our workflow is organized as follows: in Section 2, we first present an illustrative example, based on a simplified digital network, which is the basis of our motivation and to be used as a proxy in the simulation; in Section 3, we give a description of the proposed adaptive two-phase sampling design for exploring big data while minimizing costs and for estimating pertinent knowledge given the data currently available; in Section 4, we present two useful ingredients for data analysis, the first being the variance derivation and estimation used in this document, and the second being an example of the objective function when deriving the design parameter; in Section 5, we extend the notion of calibration to simultaneously adjust multiple sets of design weights in order to improve the precision of the estimates and to ensure the consistency of the estimates with constant quantities and between estimates using different or same sets of weights; finally, in Section 6, we use the simulated network to illustrate the proposed approach.

2. Illustrative Study

Digital data has augmented both the ways of understanding and the ways by which tasks are completed across an enormous array of disciplines. In this chapter, we present the context of a digital network that forms the setting for our simulation, the job shop problem, the scheduling method used in our simulation, and finally, the data generation.

2.1 Digital Network

Consider a simplified digital network which consists of two components: hosts and communication links. Each host has a unique internet protocol address allocated to it by the network. A digital network supports a number of services such as hired use of applications and data, storage servers, etc. Any given user from a population provides a service request which is sent over to the network to a certain set of hosts to perform a task on behalf of that user. Existing networks use thousands of hosts that serve hundreds of millions of users per day. The high number of requests impacts the performance of the network. The input information basically consists of the exact type of desired service and additional information including user identifiers and request times. Depending on the exact type of service required, requests can have different formats and layouts. The network analyzes the supplied information, processes it, and returns the results to the user, all the while generating a multitude of data during the process of requests. The data generated continuously over time constitutes big data.

If the network only processes a single type of request, then this network can be represented diagrammatically by means of a directed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of vertices or nodes representing hosts with two special vertices, a source V_b and a sink V_e, representing the beginning and end of the graph; and $\mathbf{E}$ is a set of edges or arrows representing distinct ordered pairs of vertices. This network can also be represented by an adjacency matrix, which is a matrix representation of exactly which nodes in the directed graph contains edges between them.

2.2 Networks of Queues

Any network is supplemented with a routing matrix which indicates the order of hosts to be visited by each service. Since each type of services has its own routing in terms of hosts, a network can be seen as a combination of S networks sharing the same hosts, where S denotes the number of available types of services. Each type s consists on a sequence of $\mathrm{O}^{(s)}$ operations or tasks, which must be performed in predetermined ordered sequence $\mathbf{O}^{(s)} = (\mathrm{o}_{(1)}, \ldots, \mathrm{o}_{(\mathrm{O}^{(s)})})$ and each operation must be processed on a given host. Each single host can process only one operation at a time and once an operation starts on a given host it must be completed on the same host without interruption. In order to increase the execution capacity of the network, a host h may represent a collection of $\mathrm{H}^{(h)}$ hosts, with $\mathrm{H}^{(h)} \geq 1$. With this representation, a host h can simultaneously perform $\mathrm{H}^{(h)}$ requests simultaneously. When all the $\mathrm{H}^{(h)}$ hosts are busy, any additional requests must wait in queue h, say $q^{(h)}$. Hence $q^{(h)}$ represents the queue is where requests waited to be served in one of the to $\mathrm{H}^{(h)}$ hosts. Since $q^{(h)}$ is attached to $\mathrm{H}^{(h)}$ hosts, a digital network can also be seen as a network of queues, in which each $q^{(h)}$ controls access to $\mathrm{H}^{(h)}$ hosts. When a single host becomes available; the first waiting request is assigned for service. If the queue processes requests in the order that they arrive; this management is termed as first come first serve. Each request k arrives at queue $q^{(h)}$ at time $a_k^{(h)}$, so that the interarrival time is $\Delta_k^{(h)} = a_k^{(h)} - a_{k-1}^{(h)}$, with $a_0^{(h)} = 0$. Once a request arrives, it waits in the queue until all previous requests have been processed. The waiting time $w_k^{(h)}$ spent in the queue $q^{(h)}$ is given by $w_k^{(h)} = \max(a_k^{(h)}, b_g^{(h)}) - a_k^{(h)}$ (Krivulin 1994), where $g = argmin\ (b_1^{(h)}, \ldots, b_{\mathrm{H}^{(h)}}^{(h)})$ and $b_1^{(h)}, \ldots, b_{\mathrm{H}^{(h)}}^{(h)}$ is the vector of times where each component g represents when single host g will next be available for request k. The response time $r_k^{(h)}$ in queue $q^{(h)}$ for request k is given by $r_k^{(h)} = w_k^{(h)} + s_{g;k}^{(h)}$, where $s_{g;k}^{(h)}$ denotes the service time in the singe host g of queue $q^{(h)}$. Once the service has been completed, the request leaves queue $q^{(h)}$. The departure time of request k is given by $d_k^{(h)} = a_k^{(h)} + r_k^{(h)}$.

2.3 Job Shop Scheduling

Scheduling is the allocation of shared hosts over time to competing requests. This scheduling task is known in the literature as “Job Shop Scheduling” or “Job Shop Problem”. The objective of the Job Shop Problem is to schedule the requests on the hosts so as to minimize some objective function such as the makespan; the time to complete all requests under constraints such as: (1) no operation can be started until the previous operation for the same request is completed, (2) a single host can only process one operation, and (3) a single host must complete processing without interruption. This optimization problem is found to be NP-hard (Muth and Thompson 1963, Garey and Johnson 1979). For our

illustration, requests are sorted in each queue $q^{(h)}$ by first arrival time and then by ascending order of their expected processing time. Let a_k denote the arrival time to the network, then the response time r_k and the departure time d_k from the network are respectively given by

$$r_k = \sum_{o \in O^{(s)}} \sum_g I_o^{(g)} r_k^{(g)},$$

and

$$d_k = a_k + r_k,$$

where the indicator variable $I_o^{(g)}$ is defined as $I_o^{(g)} = 1$ if the single host g processed operation o, and is $I_o^{(g)} = 0$ if not.

2.4 Data Generation

To create heterogeneity in our synthetic population of size N, we first generate two explanatory variables $\boldsymbol{u}_k = (u_{1;k}, u_{2;k})^T$ for each element k of the population independently from $\boldsymbol{u}_k \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})^T$, with $\boldsymbol{\mu} = (\mu_1, \mu_2)$, and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_{22} \end{bmatrix}$. We maintained the population values $\boldsymbol{u}_k$ fixed for $k = 1, \ldots, N$, and then we generated the interarrival time between requests sent by user k from an exponential distribution with parameter α_k satisfying $log(\alpha_k) = \boldsymbol{u}_{r;k}^T \boldsymbol{\beta}^{(r)}$, $\boldsymbol{u}_{r;k} = (1, u_{1;k}, u_{2;k})^T$, and $\boldsymbol{\beta}^{(r)} = (\beta_1^{(r)}, \beta_2^{(r)}, \beta_3^{(r)})^T$. For each request $r_k^{(t)}$, a service type $\mathbf{s}_k^{(t)}$ is generated from a multinomial distribution with values $\{1, 2, \ldots, S\}$, and probability $\boldsymbol{p}_k^{(t)} = (p_{1;k}^{(t)}, \ldots, p_{S;k}^{(t)})^T$, where $\mathbf{s}_k^{(t)} = (s_{1;k}^{(t)}, \ldots, s_{S;k}^{(t)})^T$, $p_{1;k}^{(t)} = (1 + \sum_{l=2}^{S} exp(\boldsymbol{u}_{s;k}^T \boldsymbol{\beta}_l^{(s)})^{-1}$, $p_{l;k} = p_{1;k} \times exp(\boldsymbol{u}_{s;k}^T \boldsymbol{\beta}_l^{(s)})$ for $l = 2, \ldots, S$, and $\boldsymbol{u}_{s;k} = (1, u_{1;k}, u_{2;k})^T$. Value of simulation parameter is provided in Table 1.

The processing time of a host is generated based on host's clock speed. A computer's clock speed, also referred to as clock rate or computer frequency, is an indicator of its performance and how rapidly a computer can process data. A higher frequency suggests better performance in common tasks. Usually, the clock speed is expressed in gigahertz (GHz) and reveals the number of instructions cycles the processor can run in a second, where one GHz is equals to E9 Hz = 1 000 000 000 Hz. For example, a 4.2GHz processor is capable of running 4.2 billion cycles per second, and a clock speed of 3.2 GHz processor executes 3.2 billion cycles per second. Sometimes, multiple instructions are completed in a single clock cycle, while in other cases, one instruction might be handled over multiple clock cycles.

Processing time of an operation i of service s in host h is generated from an exponential distribution with parameter $\Theta_{si;k}^{(h)}$, with

$$log(\Theta_{si;k}^{(h)}) = \beta_{si}^{(h)} + \gamma\, u_{3;k}^{(t)},$$

where γ is a specified parameter to account for time variation, $u_{3;k}^{(t)}$ is a random variable generated from a $N(0,1)$. The ordered operations sequence for each service type as well as values of the processing parameters are given in Tables 1 and 2. Table 2 also provides the expected processing time for each operation.

In Table 2, service 1 requires an expected execution time of $exp(.06 + .02u_{3;k}^{(t)})$ time units of host 2. The requirement of service 2 is $exp(.05 + .02u_{3;k}^{(t)})$ time units in expectation on host 1 followed by $exp(.02 + .02u_{2;k}^{(t)})$ time units in

expectation on host 3. The requirement of service 3 is $exp(.02 + .02u_{3;k}^{(t)})$ time units in expectation on host 2 followed by $exp(.04 + .02u_{3;k}^{(t)})$ time units in expectation on host 3. The requirement of service 4 is $exp(.02 + .02u_{3;k}^{(t)})$ time units in expectation on host 3 followed by $exp(.02 + .02u_{3;k}^{(t)})$ time units in expectation on host 2 followed by $exp(.02 + .02u_{3;k}^{(t)})$ time units in expectation on host 1.

Table 3 shows 4 requests as an example. The first column gives time periods, the second column gives the arrival times, the third column gives the user identifiers, the fourth column gives the service types, and the fifth column gives the request sequence. Table 6 displays the expected processing time and the hosts sequence for each request.

Table 1. Simulation Parameter

	Parameter	Value
Size	N	500
	P	5
	S	4
	H	3
heterogeneity	μ_1	**1**
	μ_2	**-1**
	σ_1	**1**
	σ_2	**1.5**
	ρ	**-.4**
Interarrival time	$\boldsymbol{\beta}^{(r)}$	(-.2, -.1 ,1)
Service Type	$\boldsymbol{\beta}_2^{(s)}$	(-.3, .4, .5)
	$\boldsymbol{\beta}_3^{(s)}$	(-.2, .3, .1)
	$\boldsymbol{\beta}_4^{(s)}$	(-.1, .3, , 2)
Processing		
	γ	.02
	$\beta_{11}^{(2)}$	.06
	$\beta_{21}^{(1)}$	.05
	$\beta_{22}^{(3)}$	.02
	$\beta_{31}^{(2)}$	.02
	$\beta_{32}^{(3)}$	.04
	$\beta_{31}^{(3)}$	.02
	$\beta_{32}^{(2)}$	.02
	$\beta_{33}^{(1)}$	.02

Table 2. Expected Host Processing Time by Operation for Each Service

Service	Operation	Host	Hosts 1	Hosts 2	Hosts 3
1	$o_{11}^{(2)}$	2		$exp(.06 + .02u_{4;k}^{(t)})$	
2	$o_{21}^{(1)}$	**1**	$exp(.05 + .02u_{4;k}^{(t)})$		
	$o_{22}^{(3)}$	**3**			$exp(.02 + .02u_{4;k}^{(t)})$
3	$o_{31}^{(2)}$	**2**		$exp(.02 + .02u_{4;k}^{(t)})$	
	$o_{32}^{(3)}$	3			$exp(.04 + .02u_{4;k}^{(t)})$
4	$o_{41}^{(3)}$	3			$exp(.02 + .02u_{4;k}^{(t)})$
	$o_{42}^{(2)}$	**2**		$exp(.02 + .02u_{4;k}^{(t)})$	
	$o_{43}^{(1)}$	**1**	$exp(.02 + .02u_{4;k}^{(t)})$		

Table 3. Example of Requests

Period(unit)	Request Time(unit)	Element Identifier	Service Type	Request Sequence
1	**1**	**1**	**3**	**1**
1	**1**	**9**	**1**	**2**
1	**1**	**11**	**2**	**3**
1	**1**	**13**	**3**	**4**
1	**1**	**14**	**1**	**5**
1	**1**	**21**	**4**	**6**

Table 4. Excepted Processing Time

Request Sequence	Hosts Expected Processing Time				Hosts Sequence		
	First	Second	Third	Total	First	Second	Third
1	**1.27**	**1.29**		**2.56**	**2**	**3**	
2	**1.32**			**1.32**	**2**		
3	**.95**	**.92**		**1.82**	**1**	**3**	
4	**1.32**	**1.35**		**2.67**	**2**	**3**	
5	**1.13**			**1.13**	**2**		
6	**1.36**	**1.36**	**1.36**	**4.07**	**3**	**2**	**1**

3. The Proposed Design

To analyse big data in discrete intervals, we divided the continuous time of the analysis period into a sequence of continuous time periods: 1, 2, etc. Suppose the limited length of duration of analysis is made up of P_{max} phases, the p^{th} being of size n_p time periods, so that the limited duration of analysis is made up of $\mathrm{I}_{max} = \sum_{p=1}^{\mathrm{P}_{max}} n_p$ time periods. This would mean that there would be a P_{max} of phases of analysis and a population total of $\mathrm{N} = \sum_{p=1}^{\mathrm{P}_{max}} \mathrm{N}_p$ during the analysis period, where N_p is the size of the finite population of interest during phase p.

Let $\boldsymbol{\chi}$ denote the prior information we have on the big data and $\boldsymbol{\psi}$ the true hidden information in the intercepted data. Under two random processes, we are primarily interested in the error-free random variable $\boldsymbol{\psi}$, knowing its probability function, the probability function of another random variable $\boldsymbol{\chi}$, together with the joint probability function $f(\boldsymbol{\psi}, \boldsymbol{\chi}; \boldsymbol{\lambda})$ of $(\boldsymbol{\chi}^T, \boldsymbol{\psi}^T)^T$ with vector parameter denoted by $\boldsymbol{\lambda}$. To choose a sample for estimation at each time period of analysis,

we primarily need to determine the design parameter $\boldsymbol{\Phi}$**,** such as the sample size and the probability of selection of each element, the solution to some objective function $O(\boldsymbol{\Phi}; \boldsymbol{\psi}, \boldsymbol{\lambda})$ that requires the target information $\boldsymbol{\psi}$ (Demnati 2019). Therefore, at each phase of data analysis, after receiving the information that the target random process has taken specific values (i.e., a new sample is selected), we update the parameter $\boldsymbol{\lambda}$ of the joint probability distribution $f(\boldsymbol{\psi}, \boldsymbol{\chi}; \boldsymbol{\lambda})$ to revise the design parameter $\boldsymbol{\Phi}$ in the course of the data analysis progression.

This section is organized as follows. Subsection 3.1 presents a straightforward description of the proposed adaptive analysis approach, while subsection 3.2 presents the proposed sampling design. Finally in subsection 3.3, we will discuss cost reduction throught the use of multiple samples.

3.1 The Steps

Our simple adaptation of the Demnati (2019)' approach for the purposes of big data analysis in discrete intervals is as follows:

a. Specify the: (1) set of possible potential parameters to be estimated, (2) population of interest for each parameter, (3) sampling frame and the sampling schemes, (4) estimator to be used for each parameter, (5) precision function for each estimator, (6) cost function of the analysis, and (7) desired precision for each estimator or the global cost of the analysis.

b. First set $\boldsymbol{\psi}_{1;k} = \boldsymbol{\chi}_k$, then for b=1, 2, … repeat continuously the following four steps until the end of data analysis

b.1 Optimization Step.

Optimize the objective function $O(\boldsymbol{\Phi}; \boldsymbol{\psi}_p, \boldsymbol{\lambda}_p)$ – that involves both the precision and cost functions parameter, i.e., determine the optimal design parameter $\boldsymbol{\Phi}$ conditional on $\boldsymbol{\psi}_p$.

b.2 Observation Step

(1) Obtain the next phase p of observations (i.e., select a new sample), and form the cumulative sample by combining all selected samples.

(2) Estimate $\boldsymbol{\theta}^{(p)}$ and $V(\widehat{\boldsymbol{\theta}}^{(p)})$ to get $\widehat{\boldsymbol{\theta}}^{(p)}$ and $\vartheta(\widehat{\boldsymbol{\theta}}^{(p)})$, where $V(\widehat{\boldsymbol{\theta}}^{(p)})$ denotes the variance of the estimator $\widehat{\boldsymbol{\theta}}^{(p)}$ based on the cumulated sample

b.3 Revision Step.

Update the vector parameter $\boldsymbol{\lambda}$ and the missing values of $\boldsymbol{\psi}$ using all available information i.e., (a) Update $\boldsymbol{\lambda}_p$ to get $\boldsymbol{\lambda}_{p+1}$ using $\mathbf{D}_{o;p}$; and (b) Impute missing values of each component ψ of $\boldsymbol{\psi}$ to get $E_{\psi}(\psi_k|\mathbf{D}_{o;p}, \boldsymbol{\lambda}_p)$, where $\mathbf{D}_{o;p}$ denotes all observed information until the end of phase p of data analysis, and E_{ψ} denotes

expectation with respect to the random process governing the component ψ. Note that $\psi_{p+1;k} = \psi_k$ when item ψ_k is observed.

b.4 Decision Step. Decide if the data analysis should stop (e.g.., $p = \mathrm{P}_{max}$). If not (e.g., $p < \mathrm{P}_{max}$), revise the specification of the design parameter as necessary and repeat the four steps (b.1 to b.4) continuously after observing some realizations of the target process.

We refer to the above four steps as the Optimization-Observation-Revision- Decision (O-O-R- D) steps. The revision step incorporates learning and prediction, while the decision step incorporates actioning.

3.2 The Sampling Design

Continuous analysis in discrete intervals may require sample rotation, samples coordination, and combination of samples through time. Using Poisson sampling in combination with a permanent random number permits the satisfaction of the above requirements and simplifies the derivation of the second-order selection probabilities $\pi_{kl}^{(st)}$, ultimately making it simpler when computing variance estimates for complex statistics such as measures of change. Here, the joint inclusion probabilities $\pi_{kl}^{(st)}$ denotes the inclusion probability of element k at time period s and element l at time period t. Under Poisson sampling, we have $\pi_{kl}^{(st)} = \pi_k^{(s)}\pi_l^{(t)}$ for $k \neq l$. From here, we must consider $\pi_{kk}^{(st)} = \pi_k^{(st)}$. If the samples selection is independent from one time period to another then $\pi_k^{(st)} = \pi_k^{(s)}\pi_k^{(t)}$ for $s \neq t$ and $\pi_k^{(tt)} = \pi_k^{(t)}$. Let $\mathrm{I}_k^{(t)} \sqsubseteq (0,1)$ denote the selection interval of element k at time t, for example $\mathrm{I}_k^{(t)} = (0, \pi_k^{(t)})$, where $\pi_k^{(t)}$ is the selection probability for element k at time t. We have $\pi_k^{(t)} = l(\mathrm{I}_k^{(t)})$ where $l(.)$ denote the length of the selection interval. The probability for an element to be selected in at least one sample over time is given by $l(\cup_t \mathrm{I}_k^{(t)})$ and the probability of selection for an element in every sample over time is given by $l(\cap_t \mathrm{I}_k^{(t)})$, where $(\cup,\cap)$ denote respectively the union and the intersection of sets. The combination of samples is just as straightforward under Poisson sampling.

We propose the use of two-phase Poisson sampling which consists here of the selection of two or more dependent Poisson samples: selection of a first large Poisson sample $\wp_0$ from a cross-sectional part of big data, where elements are consulted and selection of several second small Poisson samples $\wp_g \;\; g = 1,..,\mathrm{G}$ from $\wp_0$, which will be used for exploration at low cost. The design weights associated with the two-phase sample $\wp = (\wp_0, \wp_1, \ldots, \wp_\mathrm{G})$ are $\mathbf{w}_k = (\mathrm{w}_{0;k}, \mathrm{w}_{1;k}, \ldots, \mathrm{w}_{\mathrm{G};k})^T$, where $\mathrm{w}_{0;k} = \mathrm{w}_k(\wp_0) = a_{0;k}/\pi_{0;k}$ is the first-phase design weight, $a_{0;k} = 1_k(\wp_0)$ is the first-phase sample membership indicator variable for element k, i.e. $a_{0;k} = 1$ if $k \in \wp_0$, and $a_{0;k} = 0$ if not, $1_k(\wp_0) = 1(k \in \wp_0)$, $1()$ is the truth function, $\pi_{0;k} = E(a_{0;k})$ the first sample selection probabilities, $\mathrm{w}_{g;k} = \mathrm{w}_k(\wp_g)$, $\mathrm{w}_k(\wp_g) = \mathrm{w}_k(\wp_0)\mathrm{w}_k^{(2|1)}(\wp_g)$, $\mathrm{w}_k^{(2|1)}(\wp_g) = \mathrm{w}_k(\wp_g|k \in \wp_0) = a_{g;k}^{(2|1)}/\pi_{g;k}^{(2|1)}$ is the conditional second-phase design weight, $a_{g;k}^{(2|1)} = 1_k(\wp_g|k \in \wp_0)$ is the conditional second-phase sample $\wp_g$ membership indicator variable for element k, and $\pi_{g;k}^{(2|1)} = E(a_{g;k}^{(2|1)}|a_{0;k} = 1)$ is the subsample g selection probabilities.

Consider the general linear form given by

$$\widehat{\mathbf{U}} = \sum_k \mathbf{U}_k^T \mathbf{w}_k, \tag{3.1}$$

where $\mathbf{U}_k$ is a $(G+1) \times M$ matrix of constants and $\sum_k$ denotes the sum over the cross-sectional population elements. Under Poisson sampling at both stages, the variance of the general form (3.1) is given by

$$\mathrm{V}(\mathbf{U}) = \sum_k \mathbf{U}_k^T Cov(\mathbf{w}_k, \mathbf{w}_k) \mathbf{U}_k. \tag{3.2}$$

with $$Cov(\mathbf{w}_k, \mathbf{w}_k) = \begin{pmatrix} (1-\pi_{0;k})/\pi_{0;k} & (1-\pi_{0;k})/\pi_{0;k} & \cdots & (1-\pi_{0;k})/\pi_{0;k} \\ (1-\pi_{0;k})/\pi_{0;k} & (1-\pi_{1;k})/\pi_{1;k} & & (1-\pi_{1\mathrm{G};kl})/\pi_{1\mathrm{G};k} \\ \vdots & & & \\ (1-\pi_{0;kl})/\pi_{0;k} & (1-\pi_{\mathrm{G}1;kl})/\pi_{\mathrm{G}1;k} & & (1-\pi_{\mathrm{G};kl})/\pi_{\mathrm{G};k} \end{pmatrix},$$

with $\pi_{gg;k} = \pi_{g;k}$, $\pi_{gh;k} = \pi_{0;k}\pi_{gh;k}^{(2|1)}$ and $\pi_{gh;k}^{(2|1)} = E(a_{g;k}^{(2|1)} a_{h;k}^{(2|1)} | a_{0;k} = 1) = Pr(k \in \wp_g \cap \wp_h | a_{0;k} = 1)$,

A variance estimator of the general linear form $\widehat{\mathbf{U}}$ is given by

$$\vartheta(\mathbf{U}) = \sum_k \mathbf{U}_k^T \, cov(\mathbf{w}_k, \mathbf{w}_k) \mathbf{U}_k. \tag{3.3}$$

with $$cov(\mathbf{w}_k, \mathbf{w}_k) = diag(\mathbf{w}_k) \mathbf{C}_k diag(\mathbf{w}_k),$$

and $$\mathbf{C}_k = \begin{pmatrix} (1-\pi_{0;k}) & (1-\pi_{0;k}) & \cdots & (1-\pi_{0;k}) \\ (1-\pi_{0;k}) & (1-\pi_{1;k}) & & (1-\pi_{1\mathrm{G};k}) \\ \vdots & & & \\ (1-\pi_{0;k}) & (1-\pi_{\mathrm{G}1;k}) & & (1-\pi_{\mathrm{G};k}) \end{pmatrix}.$$

3.3 Why Not Just a Unique Sample

The answer to this question is derived from the motivation behind the traditional two-phase sampling. The traditional two-phase sampling consists of the selection of two dependent samples: selection of a first large sample $\wp_0$ in which the auxiliary variable $\boldsymbol{x}$ is measured and selection of a second small sample $\wp$ from $\wp_0$ in which the variable of interest y is measured. Two-phase sampling is appropriate when the $\boldsymbol{x}$ -values are less costly to collect than the expensive y -values. The goal of the first sample is to obtain a precise estimate related to the auxiliary variable x which can be used for sub-sampling and for estimation. For example, the two-phase ratio estimator, $\widehat{\mathrm{Y}}_R = \widehat{\mathrm{X}}_0 \widehat{Y}/\widehat{\mathrm{X}}$, is often used as an estimator of the population total $\mathrm{Y} = \sum_k y_k$, where $\widehat{\mathrm{Y}} = \sum_k \mathrm{w}_k y_k$, $\widehat{\mathrm{X}} = \sum_k \mathrm{w}_k x_k$, $\widehat{\mathrm{X}}_0 = \sum_k \mathrm{w}_{0;k} x_k$, $\mathrm{w}_{0;k}$ denotes the first-phase design weight attached to the k^{th} element, w_k denotes the design weight attached to the k^{th} element of the second-phase sample $\wp$, and $\sum_k$ denotes summation over the population elements. Two-phase sampling was first termed "double sampling" and studied by Neyman (1938) to answer the question: which sizes of the initial sample and the subsequent sample yield the most accurate estimate of the variable of interest under cost constraint. Cochran (1977, pages 327-332) and Särndal *et al*. (1992, pages 478-480) discussed the samples size determination in the case of two-phase sampling with simple random sampling at the first-phase and stratified simple random sampling at the second phase.

The double expansion estimator (Kott and Stukel, 1997), also known as "the π^* estimator" in Särndal *et al*. (1992, page 347) is the two-phase Horvitz-Thompson-type estimator of the population total. This estimator relies only on sampling design, ignoring the auxiliary information in the estimation step as the result the two-phase ratio or regression estimator, is used more as an estimator of the population total. Cochran (1977, pages 338-344) discussed the use of the ratio and regression estimators and their associated estimated variances in the case of two-phase sampling with simple random sampling in the first-phase and stratified random sampling in the second phase. Särndal *et al*. (1992, pages 343-366) extended this work for arbitrary sample designs at each phase, using the linear regression estimator to incorporate the auxiliary information. Rao and Sitter (1995) derived linearization variance estimators under the two-phase simple random sampling that takes better advantage of the available first-phase auxiliary information than the standard estimator. Demnati and Rao (2009) considered two-phase sampling in which values of the variable of interest are observed in the second-phase subsample. Values for the first-phase sample elements are mass imputed using values from an administrative file when they are available and generalized regression imputation when administrative files are not available. They studied both naïve and design-consistent estimators for a population total under the above set-up and obtained associated variance estimators. In the next chapter, we illustrate variance derivation and estimation for a calibrated estimator under the traditional two-phase sampling.

4. Two Useful Ingredients

In this section, we present the method used in this document for variance derivation and estimation. We illustrate the method with an example useful for the present work. Then, we explore an objective function used when deriving the design parameter. Finally, results of an illustration study using independent calibration estimators are presented.

4.1 Variance Derivation and Estimation Approach

We use the linearization approach of Demnati and Rao (2004, 2010) to derive variances and variance estimators. We first give a brief account of the Demnati–Rao (DR) approach. Let $\boldsymbol{d}_k = (\, d_{1;k}, d_{2;k}, \ldots, d_{p;k})^T$ be a $p\times 1$ vector of random weights and $\boldsymbol{u}_k = (u_{1;k}, \ldots, u_{p;k})^T$ be a $p \times 1$ vector of constants for $k = 1, \ldots, \mathrm{N}$, where N denotes the size of the population. Let $\widehat{\mathbf{U}} = \sum_k \mathbf{U}_k^T \boldsymbol{d}_k$ be a linear combination and, using an operator notation, let $\mathrm{V}(\boldsymbol{u})$ and $\vartheta(\boldsymbol{u})$ denote respectively the variance of $\widehat{\mathbf{U}}$ **a**nd its variance estimator. DR expressed an estimator $\widehat{\boldsymbol{\theta}}$ and its induced parameter $\boldsymbol{\theta} = E(\widehat{\boldsymbol{\theta}})$ as $\widehat{\boldsymbol{\theta}} = f(\mathbf{A}_d)$ and $\boldsymbol{\theta} = f(\mathbf{A}_{\boldsymbol{\eta}})$, where $\mathbf{A}_d$ is a $p \times \mathrm{N}$ matrix with k^{th} column $\boldsymbol{d}_k$, $\mathbf{A}_{\boldsymbol{\eta}}$ is a $p \times \mathrm{N}$ matrix with k^{th} column $\boldsymbol{\eta}_k = E(\boldsymbol{d}_k)$ and E denotes expectation under random processes involved. The DR linearization variance and variance estimator of $\widehat{\boldsymbol{\theta}} = f(\mathbf{A}_d)$ are simply given by $\mathrm{V}_{DR}(\widehat{\boldsymbol{\theta}}) = \mathrm{V}(\tilde{\boldsymbol{z}})$ and $\vartheta_{DR}(\widehat{\boldsymbol{\theta}}) = \vartheta(\boldsymbol{z})$ respectively, where $\mathrm{V}(\tilde{\boldsymbol{z}})$ is obtained from $\mathrm{V}(u)$ by replacing $\boldsymbol{u}_k$ by $\tilde{\mathbf{z}}_k = \partial f(\mathbf{A}_b)/\partial \boldsymbol{b}_k^T|_{\mathbf{A}_b = \mathbf{A}_{\boldsymbol{\eta}}}$, and $\vartheta(\boldsymbol{z})$ is obtained from $\vartheta(\boldsymbol{u})$ by replacing $\boldsymbol{u}_k$ by $z_k = \partial f(\mathbf{A}_b)/\partial \boldsymbol{b}_k^T|_{\mathbf{A}_b = \mathbf{A}_d}$, where $\mathbf{A}_b$ is a $p \times \mathrm{N}$ matrix of arbitrary real numbers with k^{th} column $\boldsymbol{b}_k = (b_{1;k}, b_{2;k}, \ldots, b_{\,p;k})^T$.

To illustrate the DR approach, we considered the population parameter $\boldsymbol{\theta}_{\mathrm{N}}$ defined as solution to "census" estimating equation of the form

$$\mathbf{S}(\boldsymbol{\theta}_{\mathrm{N}}) = \sum_k \mathbf{S}_k^T(\boldsymbol{\theta}_{\mathrm{N}}) - \boldsymbol{u}(\boldsymbol{\theta}_{\mathrm{N}}) = \mathbf{0}. \tag{4.1}$$

where $\mathbf{S}_k(\boldsymbol{\theta}_{\mathrm{N}})$ is a $p \times \mathrm{M}$ dimensional matrix-valued function of the known parameter $\boldsymbol{\theta}$ and others characteristics, and the function $\boldsymbol{u}(\boldsymbol{\theta}_{\mathrm{N}})$ allows for explicitly defined parameters. For the univariate linear and logistic regression models, $\mathbf{S}_k(\boldsymbol{\theta}) = \boldsymbol{x}_k(y_k - \mu_k(\boldsymbol{x}_k^T\boldsymbol{\theta}))$, and $\boldsymbol{u}(\boldsymbol{\theta}_{\mathrm{N}}) = \mathbf{0}$. For the special case of the finite population total $\mathrm{Y} = \sum_k y_k$, $\mathbf{S}_k(\boldsymbol{\theta}) = y_k$, $\boldsymbol{u}(\boldsymbol{\theta}_{\mathrm{N}}) = \boldsymbol{\theta}_{\mathrm{N}}$ and $\boldsymbol{\theta}_{\mathrm{N}} = \mathrm{Y}$. A design-based estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_{\mathrm{N}}$ is the solution to following weighted estimating equation

$$\widehat{\mathbf{S}}(\widehat{\boldsymbol{\theta}}) = \sum_k \mathbf{S}_k^T(\widehat{\boldsymbol{\theta}})\mathbf{w}_k - \boldsymbol{u}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{4.2}$$

Following DR approach, we write $\widehat{\boldsymbol{\theta}}$ as $f(\mathbf{A}_d)$, where $\mathbf{A}_d$ is a $\mathrm{N} \times p$ vector with k^{th} element $\boldsymbol{d}_k = \mathbf{w}_k$ and $p = \mathrm{G} + 1$. The DR variance of $\widehat{\boldsymbol{\theta}}$ is given by $Var_{DR}(\widehat{\boldsymbol{\theta}}) = \mathrm{V}(\widetilde{\boldsymbol{Z}})$, with $\widetilde{\boldsymbol{Z}}_k = \partial f(\mathbf{A}_b)/\partial b_k|_{\mathbf{A}_b = \mathbf{A}_{\boldsymbol{\eta}}}$. Taking the derivatives of $f(\mathbf{A}_b)$ and evaluating it at $\mathbf{A}_b = \mathbf{A}_{\boldsymbol{\eta}}$, we get

$$\widetilde{\boldsymbol{Z}}_k^T = \{\mathbf{J}(\boldsymbol{\theta}_{\mathrm{N}})\}^{-1}\mathbf{S}_k^T(\boldsymbol{\theta}_{\mathrm{N}}). \tag{4.3}$$

where $\mathbf{J}(\boldsymbol{\theta}_{\mathrm{N}}) = -\partial \mathbf{S}^T(\boldsymbol{\theta}_{\mathrm{N}})/\partial \boldsymbol{\theta}_{\mathrm{N}}$. The DR variance of $\widehat{\boldsymbol{\theta}}$ is given by (3.2) with $\mathbf{U}_k^T$ replaced by $\widetilde{\boldsymbol{Z}}_k^T$ given by (4.3).

Similarly, taking the derivatives of $f(\mathbf{A}_b)$ and evaluating it at $\mathbf{A}_b = \mathbf{A}_d$, we get

$$\boldsymbol{Z}_k^T = \{\hat{\mathbf{J}}(\widehat{\boldsymbol{\theta}})\}^{-1}\mathbf{S}_k^T(\widehat{\boldsymbol{\theta}}) , \tag{4.4}$$

where $\hat{\mathbf{J}}(\widehat{\boldsymbol{\theta}}) = -\partial\, \hat{\mathbf{S}}^T(\widehat{\boldsymbol{\theta}})/\partial \widehat{\boldsymbol{\theta}}$**.** The DR variance estimator of the estimator $\widehat{\boldsymbol{\theta}}$ is given by (3.3) with $\mathbf{U}_k^T$ replaced by $\boldsymbol{Z}_k^T$ given by (4.4).

4.2 Specification of the Objective Function

At phase p, the design consists on the selection of a first large Poisson sample $\wp_0^{(p)}$ from the cross-sectional population of size N_p, and selection of G second samples $\wp_g^{(p)}$ from $\wp_0^{(p)}$. We decompose the cost for each period of analysis as

$$C^{(p)} = c + C_s + C_c,$$

where c is a fixed cost. The second component C_s is associated with the sampling selection and it is given by $C_s = \sum_k a_{0;k} c_{0;k}^{(s)} + \sum_k a_{0;k} \sum_{g=1}^{G} a_{g;k}^{(2|1)} c_{g;k}^{(s)}$, where $c_{0;k}^{(s)}$ is the first-phase sampling cost for unit k and $c_{g;k}^{(s)}$ is the second-phase sampling cost for unit k. The third component C_c is associated with the process cost and it is given by $C_c, = \sum_k a_{0;k} c_{0;k}^{(c)} + \sum_k a_{0;k} \sum_{g=1}^{G} a_{g;k}^{(2|1)} c_{g;k}^{(c)}$, where $c_{0;k}^{(c)}$ is the first-phase process cost for unit k and $c_{g;k}^{(c)}$ is the second-phase process cost for unit k in subsample g.

The conditional probability that element k will be selected in in the first-phase sample is constructed as

$$log\{\pi_{0;k}/(1 - \pi_{0;k})\} = \boldsymbol{v}_{0;k}^T \boldsymbol{\Phi}_0,$$

where $\boldsymbol{v}_{0;k}$ is the vector of explanatory variable and $\boldsymbol{\Phi}_0$ is the unknown vector parameter to be determined.

Similarly, the conditional probability that element k will be selected in the second phase sample, given that the element is selected in the first-phase sample, is constructed as

$$log\{\ \pi_{g;k}^{(2|1)}/(1-\pi_{g;k}^{(2|1)})\} = \boldsymbol{v}_{g;k}^{T}\boldsymbol{\Phi}_{g},$$

where $\boldsymbol{v}_{g;k}$ is the vector of explanatory variable and $\boldsymbol{\Phi}_{g}$ is the unknown vector parameter to be determined.

To create a design at phase p i.e., to derive the design parameter $\boldsymbol{\Phi}^{(p)} = (\boldsymbol{\Phi}_0^T, \boldsymbol{\Phi}_1^T, \dots, \boldsymbol{\Phi}_{\mathrm{G}}^T)$, we minimize the conditional expected cost

$$\min_{\boldsymbol{\Phi}^{(p)}} \bar{C}^{(p)},$$

subject to constraints on Λ variances:

$$Var_{DR}(\hat{\theta}_{\kappa}^{(p)}) \leq \mathrm{V}_{\kappa}^{(p)},\ \kappa = 1,\dots,\Lambda$$

where V_{κ} are specified tolerances, and $Var_{DR}(\hat{\theta}_{\kappa})$ is the DR variance of the estimator $\hat{\theta}_{\kappa}$ for the κ^{th} parameter of interest $\kappa = 1,\dots,\Lambda$. For example, one could specify an upper limit, $\Im_{\kappa}$, on the coefficient of variation of $\hat{\theta}_{\kappa}$ so that $\mathrm{V}_{\kappa} = \{\Im_{\kappa} E(\hat{\theta}_{\kappa})\}^2$. One may repeat the optimization process with different value of $\Im_{\kappa}, \kappa = 1,\dots,\Lambda$, to obtain the desired minimum cost.

The expected cost is given by $\bar{C}^{(p)} = c + \bar{C}_s + \bar{C}_c.$, where the sampling component $\bar{C}_s$ is given by $\bar{C}_s = \sum_k \pi_{0;k} c_{0;k}^{(s)} + \sum_{g=1}^{G}\sum_k \pi_{g;k} c_{g;k}^{(s)}$. The process component $\bar{C}_c$ is given by $\bar{C}_c = \sum_k \pi_{0;k} c_{0;k}^{(c)} + \sum_{g=1}^{G}\sum_k \pi_{g;k} c_{g;k}^{(c)}$. We do not have an explicit solution, but nonlinear programming can be used to get a constrained minimum $\boldsymbol{\Phi}$.

4.3 Illustration using Independent Calibration

The two-phase ratio estimator, defined in subsection 3.3, can be viewed as a calibration estimator, $\hat{\mathrm{Y}}_R = \sum_k \tilde{w}_k y_k$, with explicit weight $\tilde{w}_k = w_k(\hat{\mathrm{X}}_0/\hat{\mathrm{X}})$ and satisfying the calibration constraint $\sum_k \tilde{w}_k x_k = \hat{\mathrm{X}}_0$. We consider first calibrated estimator $\tilde{\mathrm{Y}} = \sum_k \tilde{w}_{g;k} y_k$ of the population total $\mathrm{Y} = \sum_k y_k$ with explicit weights $\tilde{w}_{g;k} = w_{g;k}\mathrm{F}(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}_g)$ and satisfying the calibration constraint $\mathbf{S}(\hat{\boldsymbol{\lambda}}) = \sum_k w_{g;k}\mathrm{F}\big(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}\big)\boldsymbol{x}_k - \hat{\mathbf{X}}_0 = \mathbf{0}$, where $\boldsymbol{x}_k = (x_{1;k},\dots,x_{p;k})^T$. We write $\tilde{\mathrm{Y}} = f(\mathbf{A}_{\mathbf{w}})$, where $\mathbf{A}_{\mathbf{w}}$ is a $2 \times \mathrm{N}$ matrix with k^{th} column $\boldsymbol{w}_k = (w_{0;k}, w_{g;k})^T$. Taking the derivatives of $f(\mathbf{A}_{\boldsymbol{b}})$ with respect to the weight, we get

$$\frac{\partial f(\mathbf{A}_{\boldsymbol{b}})}{\partial b_{0;k}} = \mathbf{Q}_{xy}(\mathbf{A}_{\boldsymbol{b}})\frac{\partial \hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})}{\partial b_{0;k}}, \tag{4.4.a}$$

and

$$\frac{\partial f(\mathbf{A}_{\boldsymbol{b}})}{\partial b_{k}} = \mathrm{F}\big(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})\big)y_k + \mathbf{Q}_{xy}(\mathbf{A}_{\boldsymbol{b}})\frac{\partial \hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})}{\partial b_{k}}, \tag{4.4.a}$$

where $\mathbf{Q}_{xy}(\mathbf{A}_{\boldsymbol{b}}) = \sum_k b_k \acute{\mathrm{F}}(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}}))\boldsymbol{x}_k y_k$, $\acute{\mathrm{F}}(a) = \partial F(a)/\partial a$, and $\mathbf{A}_{\boldsymbol{b}}$ is a $2 \times \mathrm{N}$ matrix of constants. To get the derivatives of $\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})$, we take the derivatives of the calibration constraint.

We have $\partial\mathbf{S}(\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}}))/\partial b_{0;k} = \mathbf{Q}_{xx}(\mathbf{A}_{\boldsymbol{b}})(\partial\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})/\partial b_{0;k}) - \boldsymbol{x}_k = \mathbf{0}$, or

$$\frac{\partial \hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})}{\partial b_{0;k}} = \{\ \mathbf{Q}_{xx}(\mathbf{A}_{\boldsymbol{b}})\}^{-1}\boldsymbol{x}_k \tag{4.5.a}$$

where $\mathbf{Q}_{xx}(\mathbf{A}_{\boldsymbol{b}}) = \sum_k b_k \acute{\mathrm{F}}(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}}))\boldsymbol{x}_k\boldsymbol{x}_k^T$.

Similarly, $\partial \mathbf{S}(\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}}))/\partial b_k = \mathrm{F}\big(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})\big)\boldsymbol{x}_k + \mathbf{Q}_{xx}(\mathbf{A}_{\boldsymbol{b}})(\partial\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})/\partial b_k) = \mathbf{0}$, or

$$\frac{\partial\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})}{\partial b_k} = -\{\mathbf{Q}_{xx}(\mathbf{A}_{\boldsymbol{b}})\}^{-1}\mathrm{F}\big(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})\big)\boldsymbol{x}_k \tag{4.5.b}$$

Replacing (4.5) into (4.4) we get

$$\frac{\partial f(\mathbf{A}_{\boldsymbol{b}})}{\partial b_{0;k}} = \mathbf{Q}_{xy}(\mathbf{A}_{\boldsymbol{b}})\{\mathbf{Q}_{xx}(\mathbf{A}_{\boldsymbol{b}})\}^{-1}\boldsymbol{x}_k \tag{4.6.a}$$

$$\frac{\partial f(\mathbf{A}_{\boldsymbol{b}})}{\partial b_k} = \mathrm{F}\big(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})\big)y_k - \mathbf{Q}_{xy}(\mathbf{A}_{\boldsymbol{b}})\{\mathbf{Q}_{xx}(\mathbf{A}_{\boldsymbol{b}})\}^{-1}\mathrm{F}\big(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_{\boldsymbol{b}})\big)\boldsymbol{x}_k \tag{3.6.b}$$

Evaluating (4.6) at $\mathbf{A}_b = \mathbf{A}_\mu$, we get

$$\tilde{\boldsymbol{z}}_k = \begin{cases} \boldsymbol{x}_k^T\mathbf{B} \\ (y_k - \boldsymbol{x}_k^T\mathbf{B}) \end{cases} \tag{4.7}$$

where $\mathbf{B} = \{\sum_k \boldsymbol{x}_k\boldsymbol{x}_k^T\}^{-1}\sum_k \boldsymbol{x}_k y_k$. The DR variance of the calibrated estimator $\tilde{\mathrm{Y}} = \sum_k \tilde{w}_k y_k$ is given by $\mathrm{V}(\boldsymbol{u})$ with $\boldsymbol{u}_k = (u_{0;k}, u_k)^T$ replaced by $\tilde{\boldsymbol{z}}_k$ given by (4,7). Using (3.2), we get

$$\mathrm{V}(\boldsymbol{u}) = -\sum_k (u_{0;k} + u_k)^2 + \sum_k u_{0;k}^2/\pi_{0;k} + \sum_k\{(u_{0;k} + u_k)^2 - u_k^2\}/\pi_{g;k}. \tag{3.8}$$

While the variance of the double expansion estimator $\hat{\mathrm{Y}} = \sum_k w_{g;k} y_k$ is equal to $\mathrm{V}(\hat{\mathrm{Y}}) = -\sum_k y_k^2 + \sum_k y_k^2/\pi_{g;k}$.

Similarly, the DR variance estimator of the calibrated $\tilde{\mathrm{Y}} = \sum_k \tilde{w}_k y_k$ is simply given by $\vartheta(\boldsymbol{u})$ with $\boldsymbol{u}_k$ replaced by $\boldsymbol{z}_k$, where $\boldsymbol{z}_k = \partial f(\mathbf{A}_b)/\partial \boldsymbol{b}_k|_{\mathbf{A}_b=\mathbf{A}_\mathbf{w}}$. Evaluating (4.6) at $\mathbf{A}_b = \mathbf{A}_\mathbf{w}$, we get respectively

$$\boldsymbol{z}_k = \begin{cases} \boldsymbol{x}_k^T\hat{\mathbf{B}} \\ \mathrm{F}(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})\ (y_k - \boldsymbol{x}_k^T\hat{\mathbf{B}}) \end{cases} \tag{4.9}$$

where $\hat{\mathbf{B}} = \{\sum_k w_k \acute{\mathrm{F}}(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})\boldsymbol{x}_k\boldsymbol{x}_k^T\}^{-1}\sum_k w_k \acute{\mathrm{F}}(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}})\boldsymbol{x}_k y_k$.

We first generate two explanatory variables $\boldsymbol{u}_k = (u_{1;k}, u_{2;k})^T$ for each element k of the population independently from $\boldsymbol{u}_k \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})^T$, with $\boldsymbol{\mu} = (\mu_1, \mu_2)$, and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_{22} \end{bmatrix}$. We set $\boldsymbol{\mu} = (1, -1)$, $\sigma_1 = \sigma_2 = 5$, and $\rho = .4$. We maintained the population values $\boldsymbol{u}_k$ fixed for $k = 1, \ldots, \mathrm{N}$, and then we generated $y_{1;k}$ from the Bernoulli distribution with success probability satisfying $\mathrm{logit}(p_k) = \boldsymbol{v}_k^T\boldsymbol{\beta}_1$ with $\boldsymbol{\beta}_1 = (-.3, .4, .6)$ and $\boldsymbol{v}_k = (1, x_{1;k}, x_{2;k})^T$, and $y_{2;k}$ from the normal distribution with mean $\mu_k = \boldsymbol{v}_k^T\boldsymbol{\beta}_2$ with $\boldsymbol{\beta}_2 = (20,2,5)$ and variance equals to 25. Two parameters are of interest, $\theta_{\mathrm{N}}^{(i)} = \sum_k y_{i;k}$ $i \in \{1,2\}$ Three scenarios are considered for the estimation of $\theta_{\mathrm{N}}^{(i)}$: 1) $\hat{\theta}_1^{(i)} = \sum_k w_{0;k} y_{i;k}$ using the first-phase sample only; 2) $\hat{\theta}_2^{(1)} = \sum_k w_{g;k} y_{1;k}$, and $\hat{\theta}_2^{(2)} = \sum_k w_{h;k} y_{2;k}$ using the double expansion estimator and the two-phase sample g and h with $g \neq h$; and, 3) $\tilde{\theta}_3^{(1)} = \sum_k \tilde{w}_{g;k} y_{1;k}$ and $\tilde{\theta}_3^{(2)} = \sum_k \tilde{w}_{h;k} y_{2;k}$ with explicit weight $\tilde{w}_{g;k} = w_{g;k}\mathrm{F}(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}_g)$ and $\tilde{w}_{h;k} = w_{h;k}\mathrm{F}(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}_h)$, respectively and satisfying the calibration constraint $\mathbf{S}(\hat{\boldsymbol{\lambda}}_g)) = \sum_k w_{g;k}\mathrm{F}\big(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}_g)\big)\boldsymbol{x}_k - \hat{\mathbf{X}}_0 = \mathbf{0}$ and $\mathbf{S}(\hat{\boldsymbol{\lambda}}_h)) = \sum_k w_{h;k}\mathrm{F}\big(\boldsymbol{x}_k^T\hat{\boldsymbol{\lambda}}_h)\big)\boldsymbol{x}_k - \hat{\mathbf{X}}_0 = \mathbf{0}$, where $\boldsymbol{x}_k = (1, x_{1;k}, x_{2;k})^T$ and $g \neq h$. We first write each estimator as $f(\mathbf{A}_\mathbf{w})$, where $\mathbf{A}_\mathbf{w}$ is a $2 \times \mathrm{N}$ matrix with k^{th} column $\boldsymbol{w}_k = (w_{0;k}, w_{g;k})^T$, then we take the derivatives of $f(\mathbf{A}_{\boldsymbol{b}})$ with respect to the weight, and we evaluate it at $\mathbf{A}_b = \mathbf{A}_\mu$, to get,

for scenario 1, $\tilde{\boldsymbol{z}}_{i;k} = \begin{cases} y_{i;k} \\ 0 \end{cases}$

for scenario 2, $\tilde{\boldsymbol{z}}_{i;k} = \begin{cases} 0 \\ y_{i;k} \end{cases}$

for scenario 3, $$\tilde{\boldsymbol{z}}_k = \begin{cases} \boldsymbol{x}_k^T \mathbf{B}_i \\ (y_{i;k} - \boldsymbol{x}_k^T \mathbf{B}_i) \end{cases} \equiv \begin{cases} \tilde{z}_{i0;k} \\ \tilde{z}_{ig;k} \end{cases}$$

where $\mathbf{B}_i = \{\sum_k \boldsymbol{x}_k \boldsymbol{x}_k^T\}^{-1} \sum_k \boldsymbol{x}_k y_{i;k}$.

Under scenario 1, the optimization problem consists on the determination $\boldsymbol{\Phi} = (\Phi_0^{(1)}, \Phi_0^{(2)}, \Phi_0^{(3)})^T$ such that the expected cost $\bar{C} = \sum_k \pi_{0;k} c_k$, is minimized subject to constraints on the variance

$$-\sum_k y_{1;k}^2 + \sum_k y_{1;k}^2/\pi_{0;k} \leq \mathrm{V}_1,$$

and $$-\sum_k y_{2;k}^2 + \sum_k y_{2;k}^2/\pi_{0;k} \leq \mathrm{V}_2.$$

Under scenario 2, the optimization problem consists on the determination $\boldsymbol{\Phi} = (\Phi_0^{(1)}, \Phi_0^{(2)}, \Phi_0^{(3)}, \Phi_g^{(1)}, \Phi_g^{(2)}, \Phi_g^{(3)}, \Phi_h^{(1)}, \Phi_h^{(2)}, \Phi_h^{(3)})^T$ such that the expected cost $\bar{C} = \sum_k (\pi_{0;k} c_{0;k} + \pi_{g;k} c_{g;;k} + \pi_{h;k} c_{h;k})$, with $c_{0;k} + c_{g;;k} + c_{h;k} = c_k$ and $g \neq h$, is minimized subject to constraints on the variance

$$-\sum_k y_{1;k}^2 + \sum_k y_{1;k}^2/\pi_{g;k} \leq \mathrm{V}_1,$$

and $$-\sum_k y_{2;k}^2 + \sum_k y_{2;k}^2/\pi_{h;k} \leq \mathrm{V}_2.$$

Under scenario 3, the optimization problem consists on the determination $\boldsymbol{\Phi} = (\Phi_0^{(1)}, \Phi_0^{(2)}, \Phi_0^{(3)}, \Phi_g^{(1)}, \Phi_g^{(2)}, \Phi_g^{(3)}, \Phi_h^{(1)}, \Phi_h^{(2)}, \Phi_h^{(3)})^T$ such that the expected cost $\bar{C} = \sum_k (\pi_{0;k} c_{0;k} + \pi_{g;k} c_{g;;k} + \pi_{h;k} c_{h;k})$, with $c_{0;k} + c_{g;;k} + c_{h;k} = c_k$ and $g \neq h$, is minimized subject to constraints on the variance

$$-\sum_k (\tilde{z}_{10;k} + \tilde{z}_{1g;k})^2 + \sum_k \{(\tilde{z}_{10;k} + \tilde{z}_{1g;k})^2 - \tilde{z}_{1g;k}^2\}/\pi_{0;k} + \sum_k \tilde{z}_{1g;k}^2/\pi_{g;k} \leq \mathrm{V}_1,$$

and $$-\sum_k (\tilde{z}_{20;k} + \tilde{z}_{2h;k})^2 + \sum_k \{(\tilde{z}_{20;k} + \tilde{z}_{2h;k})^2 - \tilde{z}_{2h;k}^2\}/\pi_{0;k} + \sum_k \tilde{z}_{2h;k}^2/\pi_{h;k} \leq \mathrm{V}_2.$$

Here $\mathrm{V}_i = (\mathrm{cv} \times \sum_k y_{i;k})^2$ with $cv = .05$, and $\pi_{g;k} = \pi_{0;k} \pi_{g;k}^{(2|1)}$. The derivatives of $\pi_{0;k}$ and $\pi_{g;k}$ are respectively $\dot{\boldsymbol{\pi}}_{0;k} = \boldsymbol{x}_{0;k} \pi_{0;k}(1 - \pi_{0;k})$ and $\dot{\boldsymbol{\pi}}_{g;k} = \boldsymbol{x}_{0;k} \pi_{g;k}(1 - \pi_{0;k}) + \boldsymbol{x}_{g;;k} \pi_{g;k}(1 - \pi_{g;k}^{(2|1)})$. These derivatives are useful for the optimization procedure. We set $(c_{0;k}, c_{1;;k}, c_{2;k}) = (1,10,20)$. Table 5, displays results of the illustration. It is clear from Table 5 that when the costs and characteristics differ from one variable to another, the use of different sub-samples makes it possible to reduce the total cost.

Table 5: Result of the Sample Size Determination Using $(c_{0;k}, c_{1;;k}, c_{2;k}) = (1,10,20)$

Strategy	Calibration	Expected Total Cost	Expected Sample Size		
			$\wp_0$	$\wp_1$	$\wp_2$
$\wp_0$	No	7 281	108		
$\wp_0, \wp_1, \wp_2$	No	2 381	176	77	72
$\wp_0, \wp_1, \wp_2$	Yes	1 589	160	53	45

5. Multiple Weights Calibration

Poisson sampling, an easy sampling scheme, simplifies samples combination and variance computation. It is also a flexible way to incorporate auxiliary information into the design at the sampling stage. However, Poisson sampling presents a drawback because the sample size is random which causes a very large variance of any Horvitz-Thompson

(HT) estimator of a population total. For example, the HT estimator of the total under Bernoulli sampling has a much greater variance than under simple random sampling. Fortunately, the use of calibrated estimators eliminates this drawback almost entirely. Calibration is also used to incorporate auxiliary information into the design at the estimation step, and to ensure consistency with fixed quantities such as finite population known totals. In the context of this work of multiple sets of weights, it is desirable to maintain consistency between estimates from multiple sets of weights coming from different samples selected from the same population.

5.1 A General Class of Regression Calibration Weights

Let $\boldsymbol{d}_k$ be the $p \times 1$ vector of random weights, $\boldsymbol{\eta}_k = E(\boldsymbol{d}_k)$, $\mathbf{Y}_k$ be a $p \times \mathrm{M}$ design matrix related to the statistics of interest, $\mathbf{X}_{d;k}$ be a $p \times q$ design matrix related to the auxiliary variables, and $\mathbf{X}_{\eta;k}$ be a $p \times q$ design matrix related to the expected values. Let the estimator be $\widehat{\boldsymbol{\theta}} = \sum_k \mathbf{Y}_k^T \boldsymbol{d}_k$ and zero-mean linear equation be $\widehat{\mathbf{X}}_\circ = \sum_k (\mathbf{X}_{d;k}^T \boldsymbol{d}_k - \mathbf{X}_{\eta;k}^T \boldsymbol{\eta}_k)$. For example consider the traditional two-phase sampling design with $\boldsymbol{d}_k = (\mathrm{w}_{0;k}, \mathrm{w}_k)^T$. If $\widehat{\mathbf{Y}}$ denotes the estimator $\widehat{\mathbf{Y}} = (\widehat{\mathrm{Y}}_0, \widehat{\mathrm{Y}})^T$ of the population total $\mathbf{Y} = \sum_k \boldsymbol{y}_k$ with $\boldsymbol{y}_k = (y_{0;k}, y_k)^T$, then $\boldsymbol{\theta} = \sum_k \mathbf{Y}_k^T \boldsymbol{\eta}_k$, and $\widehat{\boldsymbol{\theta}} = \sum_k \mathbf{Y}_k^T \boldsymbol{d}_k$, with $\mathbf{Y}_k = diag(\boldsymbol{y}_k)$, where $\boldsymbol{\eta}_k = E(\boldsymbol{d}_k) = (1,1)^T$, $\mathrm{G} = 1$, and any zero-mean linear equation can be represented by $\widehat{\mathbf{X}}_\circ = \sum_k (\mathbf{X}_{d;k}^T \boldsymbol{d}_k - \mathbf{X}_{\eta;k}^T \boldsymbol{\eta}_k)$. For the zero-mean linear equation $\hat{X}_0 - \hat{X}$, $q = 1$, $\mathbf{X}_{d;k} = x_k(1,-1)^T$ and $\mathbf{X}_{\eta;k} = (0,0)^T$, while for the zero-mean linear equation $(\hat{X}_0 - X, \hat{X} - X)^T$, $q = 2$, $\mathbf{X}_{d;k} = x_k \mathbf{I}_2$ and $\mathbf{X}_{\eta;k} = x_k \mathbf{I}_2$, where $\mathbf{I}_2$ is the identity matrix. Here, $\widehat{\mathrm{Y}}_0 = \sum_k w_{0;k} y_{0;k}$, $\widehat{\mathrm{Y}} = \sum_k w_k y_k$ and w_k are the design weights of the two-phase sample.

We define the regression calibrated estimator of the M × 1 linear quantity $\boldsymbol{\theta} = \sum_k \mathbf{Y}_k^T \boldsymbol{\eta}_k$ as

$$\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} - \widehat{\mathbf{B}}^T \widehat{\mathbf{X}}_\circ, \tag{5.1}$$

with $$\widehat{\mathbf{B}} = \widehat{\mathbf{Q}}_{X_d X_d}^{-1} \widehat{\mathbf{Q}}_{X_{d;}Y},$$

where $\widehat{\boldsymbol{\theta}} = \sum_k \mathbf{Y}_k^T \boldsymbol{d}_k$, $\widehat{\mathbf{Q}}_{AW} = \sum_k \sum_l \mathbf{A}_k^T \, cov(\boldsymbol{d}_k, \boldsymbol{d}_l) \mathbf{W}_l$, and $cov(\boldsymbol{d}_k, \boldsymbol{d}_l) = diag(\boldsymbol{d}_k) \mathbf{C}_{kl} diag(\boldsymbol{d}_l)$. We may write the ij^{th} element of $\mathbf{C}_{kl}$ as $c_{ij;kl} = Cov(d_{i;k}, d_{j;l}) / E(d_{i;k} d_{j;l})$ with $Cov(d_{i;k}, d_{j;l}) = E(d_{i;k} d_{j;l}) - E(d_{i;k}) E(d_{j;l})$. The regression calibrated estimator (5.1) can also be written in term of the g-factors

$$\widetilde{\boldsymbol{\theta}} = \sum_k \mathbf{Y}_k^T diag(\boldsymbol{d}_k) \mathbf{G}_k,$$

with $$\mathbf{G}_k = \mathbf{1}_p - \sum_l \mathbf{C}_{kl} diag(\boldsymbol{d}_l) \mathbf{X}_{d;l} \widehat{\mathbf{Q}}_{X_d X_d}^{-1} \widehat{\mathbf{X}}_\circ, \tag{5.2}$$

where $\mathbf{1}_p$ is the $p \times 1$ vector of ones. Note that the generalized regression (Särndal et al. 1989) and the optimal linear regression (Montanari 1998) are special cases of (5.1).

Illustrations under Traditional Two-phase sampling

For the zero-mean linear equation $\hat{X}^{(1)} - X$, we have $q = 1$, $\mathbf{X}_{d;k} = x_k(1,0)^T$, $\mathbf{X}_{\eta;k} = x_k(1,0)^T$, and

$$\mathbf{G}_k = \binom{1}{1} - \binom{\sum_l w_{0;l} x_l (1 - \omega_{0;kl})}{\sum_l w_{0;l} x_l (1 - \omega_{0;kl})} (\hat{X}^{(1)} - X) / \vartheta(\hat{X}^{(1)}).$$

If $\boldsymbol{y}_k = (x_k, 0)^T$ then $\tilde{\theta} = \mathrm{X}$, the population total. If $\boldsymbol{y}_k = (0, x_k)^T$ then $\tilde{\theta} = \hat{X} - \vartheta(\hat{X}^{(1)}, \hat{X})(\hat{X}^{(1)} - X) / \vartheta(\hat{X}^{(1)})$, the "optimal" estimator from the subsample given the first-phase sample estimate.

For the zero-mean linear equation $\hat{X}_1^{(1)} - \hat{X}$, we have $q = 1$, $\mathbf{X}_{d;k} = x_k(1,-1)^T$, $\mathbf{X}_{\eta;k} = (0,0)^T$, and

$$\mathbf{G}_k = \binom{1}{1} - \begin{pmatrix} \sum_l (w_{0;l} - w_l)x_l(1-\omega_{0;kl}) \\ \sum_l \{w_{0;l}(1-w_{0;kl}) - w_k d_{2;l}(1-\omega_{kl})\}x_l \end{pmatrix} (\hat{X}^{(1)} - \hat{X})/\vartheta(\hat{X}^{(1)} - \hat{X}).$$

If $\boldsymbol{y}_k = (x_k, 0)^T$ then $\tilde{\theta} = \hat{X}^{(1)} - \{\vartheta(\hat{X}^{(1)}) - \vartheta(\hat{X}^{(1)}, \hat{X})(\hat{X}^{(1)} - \hat{X})/\vartheta(\hat{X}^{(1)} - \hat{X})$. If $\boldsymbol{y}_k = (0, x_k)^T$ then, $\tilde{\theta} = \hat{X} - \{\vartheta(\hat{X}) - \vartheta(\hat{X}^{(1)}, \hat{X})(\hat{X}^{(1)} - \hat{X})/\vartheta(\hat{X}^{(1)} - \hat{X})$.

Finally, for example for the zero-mean linear equation $\hat{\mathrm{X}} - \mathrm{X}$, we have $q = 1$, $\mathbf{X}_{d;k} = x_k(0,1)^T$, $\mathbf{X}_{\eta;k} = (0,1)^T$, and

$$\mathbf{G}_k = \binom{1}{1} - \begin{pmatrix} \sum_l w_l x_l(1-\omega_{kl}) \\ \sum_l w_l x_l(1-\omega_{kl}) \end{pmatrix} (\hat{X} - X)/\vartheta(\hat{X}).$$

If $\boldsymbol{y}_k = (x_k, 0)^T$ then $\tilde{\theta} = \hat{X}^{(1)} - \vartheta(\hat{X}^{(1)}, \hat{X})(\hat{X} - X)/\vartheta(\hat{X})$, the "optimal" estimator from the first-phase sample given the subsample estimate. If $\boldsymbol{y}_k = (0, x_k)^T$ then $\tilde{\theta} = \mathrm{X}$, the population total.

5.2 A General Class of Calibrated Estimators

The calibration weights associated with the g-factors (5.2) may not be always nonnegative. To get around this difficulty in the univariate weight case, generalized raking ratio weights are often used. These weights are always nonnegative, but the method can lead to some extreme weights. The univariate generalized raking weights can also be extended under multiple design weights. We modify the regression calibrated estimator (5.1) as follows

$$\tilde{\boldsymbol{\theta}} = \sum_k \mathbf{Y}_k^T diag(\boldsymbol{d}_k)\mathbf{F}(\mathbf{X}_{d;k}\hat{\boldsymbol{\lambda}}), \tag{5.3.a}$$

with $$\mathbf{F}(\mathbf{X}_{d;k}\boldsymbol{\lambda}) = (\mathrm{F}(\mathbf{X}_{1d;k}\boldsymbol{\lambda}), \ldots, \mathrm{F}(\mathbf{X}_{pd;k}\boldsymbol{\lambda})),$$

where $\mathbf{X}_{id;k}^T$ is the i^{th} rows of the matrix $\mathbf{X}_{d;k}$, the $q \times 1$ vector estimator $\hat{\boldsymbol{\lambda}}$ is the solution to the calibration equation

$$\mathbf{S}(\hat{\boldsymbol{\lambda}}) = \sum_k (\mathbf{X}_{d;k}^T diag(\boldsymbol{d}_k)\mathbf{F}(\mathbf{X}_{d;k}\hat{\boldsymbol{\lambda}}) - \mathbf{X}_{\eta;k}^T \boldsymbol{\eta}_k) = \mathbf{0}. \tag{5.3.b}$$

5.3 Variance Estimation for the General Calibrated Estimators

We used the general calibrated weights given by (5.3). We have

$$\partial\,(b_{g;k}\mathrm{F}(\mathbf{X}_{gd;k}^T\boldsymbol{\lambda}(\mathbf{A}_b)))/\partial \boldsymbol{b}_{g;k} = \mathrm{F}(\mathbf{X}_{gd;k}^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_b) + b_{g;k}\dot{\mathbf{F}}(\mathbf{X}_{gd;k}^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_b))(\partial\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)/\partial \boldsymbol{b}_{g;k}), \tag{5.4.a}$$

and for $j \neq g$ or $l \neq k$

$$\partial\,(b_{j;l}\mathrm{F}(\mathbf{X}_{jd;k}^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_b))/\partial \boldsymbol{b}_{g;k} = b_{j;l}\dot{\mathbf{F}}(\mathbf{X}_{jd;k}^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_b))(\partial\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)/\partial \boldsymbol{b}_{g;k}), \tag{5.4.b}$$

where $\dot{\mathbf{F}}(.) = \partial\mathbf{F}(.)/\partial\boldsymbol{\lambda}^T$

To evaluate $(\partial\boldsymbol{\lambda}(\mathbf{A}_b)/\partial b_{i;k})$ we take the derivatives of the calibration EE (5.3.b) with respect to $b_{i;k}$. This gives

$$\mathbf{X}_{gd;k}^T\mathrm{F}(\mathbf{X}_{gd;k}^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)) + \sum_k (\mathbf{X}_{d;k}^T diag(\boldsymbol{b}_k)\dot{\mathbf{F}}(\mathbf{X}_{d;k}^T\boldsymbol{\lambda}(\mathbf{A}_b))\partial\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)/\partial b_{g;k} = \mathbf{0},$$

or $$\partial\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)/\partial b_{g;k} = -\widehat{\mathbf{Q}}_\lambda^{-1}(\mathbf{A}_b)\mathbf{X}_{gd;k}\mathrm{F}(\mathbf{X}_{gd;k}^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)), \tag{5.5}$$

where $\widehat{\mathbf{Q}}_\lambda(\mathbf{A}_b) = \sum_k (\mathbf{X}_{d;k}^T diag(\boldsymbol{b}_k)\acute{\mathbf{F}}(\mathbf{X}_{d;k}^T\hat{\boldsymbol{\lambda}}(\mathbf{A}_b))$.

Replacing (5.5) into (5.4), we get

$$\partial\,(b_{\,g;k}\mathrm{F}(\mathbf{X}_{d;k}^{T}\boldsymbol{\lambda}(\mathbf{A}_b)))/\partial\boldsymbol{b}_{\,g;k} = \mathrm{F}(\mathbf{X}_{gd;k}^{T}\hat{\boldsymbol{\lambda}}(\mathbf{A}_b) - b_{\,g;k}\dot{\mathbf{F}}(\mathbf{X}_{gd;k}^{T}\hat{\boldsymbol{\lambda}}(\mathbf{A}_b))\widehat{\mathbf{Q}}_{\lambda}^{-1}(\mathbf{A}_b)\mathbf{X}_{gd;k}\mathrm{F}(\mathbf{X}_{gd;k}^{T}\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)), \quad (5.6.a)$$

and for $j \neq g$ or $l \neq k$

$$\partial\,(b_{\,j;l}\mathrm{F}(\mathbf{X}_{jd;k}^{T}\hat{\boldsymbol{\lambda}}(\mathbf{A}_b))/\partial\boldsymbol{b}_{\,g\,;k} = -b_{\,j;l}\dot{\mathbf{F}}(\mathbf{X}_{jd;k}^{T}\hat{\boldsymbol{\lambda}}(\mathbf{A}_b))\widehat{\mathbf{Q}}_{\lambda}^{-1}(\mathbf{A}_b)\mathbf{X}_{gd;k}\mathrm{F}(\mathbf{X}_{gd;k}^{T}\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)), \quad (5.6.b)$$

For the linear calibrated estimator (3.1), $f(\mathbf{A}_d) = \sum_k \mathbf{U}_k^T\mathbf{w}_k = \sum_k\sum_g y_{g;k}\widetilde{\mathrm{w}}_{g;k}$, we get

$$\partial f(\mathbf{A}_b)/\partial\boldsymbol{b}_{g;k} = \mathrm{F}(\mathbf{X}_{gd;k}^{T}\hat{\boldsymbol{\lambda}}(\mathbf{A}_b)(y_{g\,;k} - \mathbf{B}_{\boldsymbol{\lambda}}(\mathbf{A}_b)\mathbf{X}_{gd;k}\}, \quad (5.7)$$

where $\mathbf{B}_{\boldsymbol{\lambda}}(\mathbf{A}_b) = \sum_k\sum_g y_{g\,;k}\{b_{\,g;k}\dot{\mathbf{F}}(\mathbf{X}_{gd;k}^{T}\hat{\boldsymbol{\lambda}}(\mathbf{A}_b))\widehat{\mathbf{Q}}_{\lambda}^{-1}(\mathbf{A}_b)$.

The DR variance estimator of $f(\mathbf{A}_d)$ is given by $\vartheta(\,\mathbf{z})$, where $\mathbf{z}_{g;k} = \partial f(\mathbf{A}_b)/\partial\boldsymbol{b}_{g;k}|_{A_b=A_d}$. Evaluating (5.6) at $\mathbf{A}_b = \mathbf{A}_{\boldsymbol{d}}$, we get

$$\boldsymbol{z}_{g;k} = \mathrm{F}(\mathbf{X}_{gd;k}^{T}\hat{\boldsymbol{\lambda}}(y_{g;k} - \widehat{\mathbf{B}}_{\boldsymbol{\lambda}}\mathbf{X}_{gd;k}\},$$

where $\widehat{\mathbf{B}}_{\lambda} = \sum_k\sum_g y_{g;k}d_{\,g;k}\dot{\mathbf{F}}(\mathbf{X}_{id;k}^{T}\hat{\boldsymbol{\lambda}})\widehat{\mathbf{Q}}_{\lambda}^{-1}$ and $\widehat{\mathbf{Q}}_{\lambda} = \sum_k(\mathbf{X}_{d;k}^{T}diag(\boldsymbol{d}_k)\acute{\mathbf{F}}(\mathbf{X}_{d;k}^{T}\hat{\boldsymbol{\lambda}})$.

Concluding Remarks

We formulated an adaptive approach for digital data exploration and pertinent knowledge extraction. We also proposed a simple sampling design that permits desirable properties such as samples combination and coordination over time periods, as well as variance estimation for complex estimators. The adaptive aspect of the proposed approach and the use of several sub-samples makes it possible to both adequately target the parameters of interest over time and reduce total costs, while all sampled elements are used by easily combining them. Calibration of multiple sets of weights is also completed to ensure the consistency of the estimates with constant quantities as well as between estimates using different or same sets of weights. Further simulations will be completed at a later time.

Acknowledgement

This work is dedicated to Dr. Patricia Mouaikel and her staff for the perfect combination of responsibility and care she offers to her patients. Over the past several decades, we have shared in many fruitful conversations and the commitment she shows to her vocation continues to inspire me in my own.

References

Cochran, W.G. 1977. *Sampling Techniques*, 3rd edn. New York: John Wiley.

Demnati, A. 2019. Responsive Design – Side Effect Reduction of Prior and Processed Information on Survey Design. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. Denver, Colorado, USA, July 27 - August 1.

Demnati, A. and Rao, J.N.K. 2004. Linearization Variance Estimators for Survey Data (with discussion). *Survey Methodology*, 30, 17-34.

Demnati, A. and Rao, J.N.K. 2009. Linearization Variance Estimation and Allocation for Two-phase Sampling under Mass Imputation. *Federal Committee on Statistical Methodology Research Conference,* American Statistical Association

Demnati, A. and Rao, J.N.K. 2010. Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, 36, 193-199.

Garey, M. and Johnson, D. 1979. *Computers and intractability*, Freeman San Francisco, CA.

Kott. P.S., and Stukel, D.M., 1997. Can the jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-90.

Krivulin, N. K, 1994. A recursive equations-based representation for the G/G/M queue, *Applied Mathematics Letters*, 7, 73-77.

Montanari, G. E. 1998. On regression estimation of finite population means. Survey methodology, 24, 69-77.

Muth, J. F. and Thompson, G. L. 1963. *Industrial Scheduling*. Prentice-Hall, Engle-wood Cliffs, N.J.

Neyman, J. 1938. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

Rao, J.N.K., and Sitter, R.R. 1995. Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Särndal, C.-E., Swensson, B.E., and Wretman, J.H. 1989. The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

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