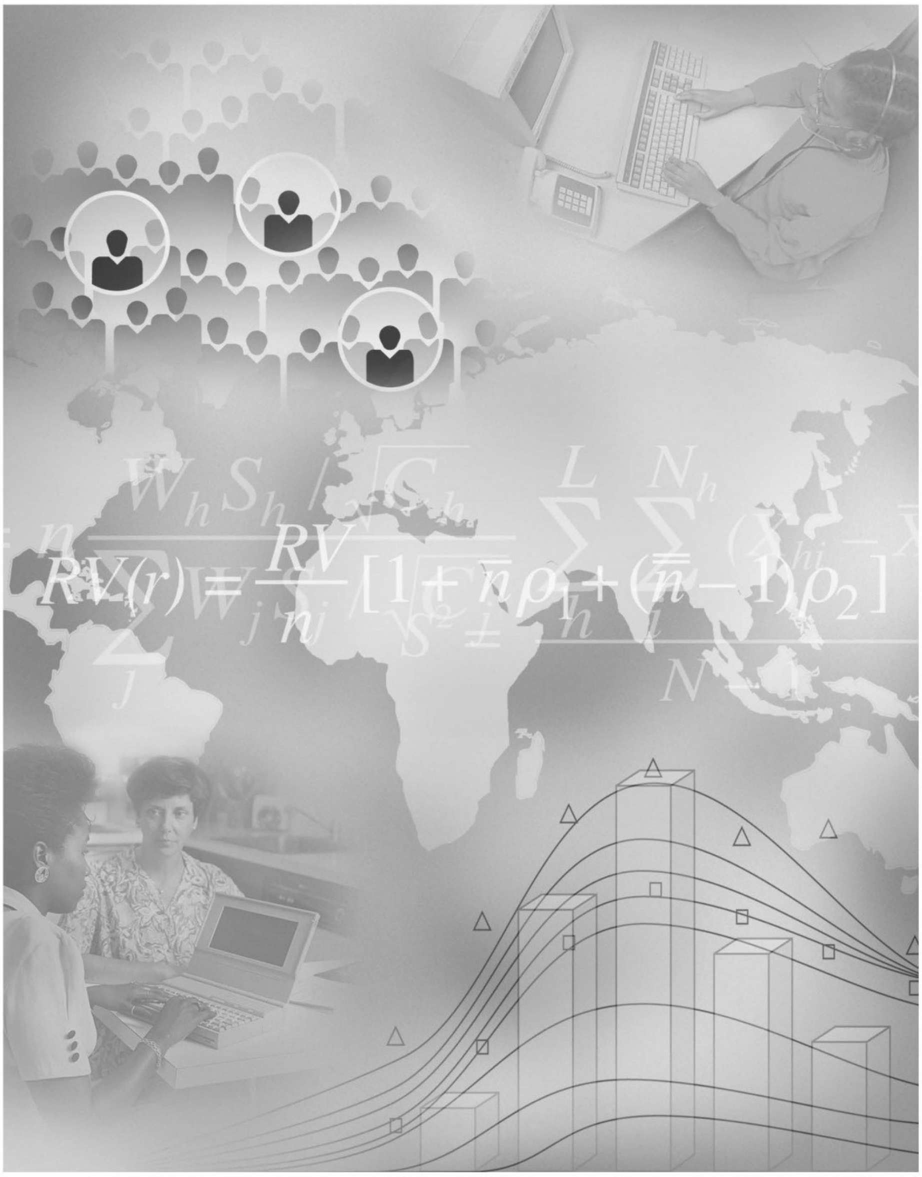
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

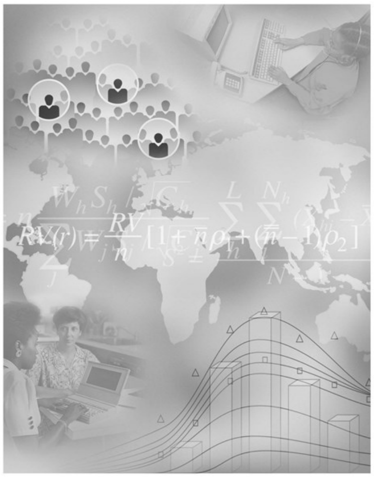
No. 83 January 2021


$RV(r) = \frac{RV}{n}[1+\bar{n}\rho_1+(\bar{\bar{n}}-1)\rho_2]$

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

the Survey Statistician
The Newsletter of the International Association of Survey Statisticians



The Survey Statistician No. 83, January 2021

Editors:
Danutė Krapavickaitė (*Vilnius Gediminas Technical University, Lithuania*) and Eric Rancourt (*Statistics Canada*)

Section Editors:

Peter Wright	Country Reports
Ton de Waal	Ask the Experts
James Chipperfield	New and Emerging Methods

Production and Circulation:
Mārtiņš Liberts and Liliāna Roze (*Central Statistical Bureau of Latvia*), Melini Cooper and Harry Raymond (*Australian Bureau of Statistics*)

The Survey Statistician is published twice a year by the International Association of Survey Statisticians and distributed to all its members. The Survey Statistician is also available on the IASS website at http://isi-iass.org/home/services/the-survey-statistician/

Enquiries for membership in the Association or change of address for current members should be addressed to:

IASS Secretariat Membership Officer
Margaret de Ruiter-Molloy
International Statistical Institute
P.O. Box 24070, 2490 AB the Hague
The Netherlands

Comments on the contents or suggestions for articles in the Survey Statistician should be sent via e-mail to the editors:
Danutė Krapavickaitė (danute.krapavickaite@vilniustech.lt) or Eric Rancourt (eric.rancourt@canada.ca).

ISSN 2521-991X

In this Issue

3 Letter from the Editors

4 Letter from the President

6 Report from the Scientific Secretary

8 News and Announcements
- ISI World Statistics Congress 2021 goes VIRTUAL!
- ISI World Statistics Congress 2023: OTTAWA is chosen!
- Announcement: New editor of Survey Methodology
- Analysis of survey data using R short course
- A tribute to Ken Brewer

11 Ask the Experts
- Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? By Jean-Francois Beaumont, and J. N. K. Rao (***reviewed paper***)
- Official Statistics at the Crossroads: Data Quality and Access in an Era of Heightened Privacy Risk, by John M. Abowd (***reviewed paper***)

27 New and Emerging Methods
- Graph sampling: An introduction, by Li-Chun Zhang (***reviewed paper***)

38 Country Reports
- Argentina
- Australia
- Canada
- Fiji
- New Zealand

43 Upcoming Conferences and Workshops

44 In Other Journals

51 Welcome New Members!

52 IASS Executive Committee Members

53 Institutional Members

Letter from the Editors

Dear Readers, happy New Year 2021!

Last year we wished you health but unfortunately, our wish did not materialize for the whole world. Now we would like to wish you to resist to the disasters and to continue your activities in the best possible adapted way. In this vein, the current issue of TSS presents papers on ways to adapt the statistical work to changes in the world: Dealing with rising non-probability samples from large amounts of data and their integration with probability samples; finding a balance between high quality statistical results and data confidentiality; performing graph sampling using graphs which may reflect current spread of the virus.

The country reports present the work of statistical offices under pandemic conditions: Population census in Argentina; visualization of statistical results in Australia; use of machine learning techniques for statistical production in Canada; recent and upcoming surveys in Fiji; and a Covid-19 data portal in New Zealand. The issue presents information about the forthcoming conferences including the 63rd World Statistics Congress online and the contents of journals of interest to survey statisticians.

We would like to thank everyone who has provided inputs to this issue, all authors of the articles, and all people who worked in the preparation of the current issue. James Chipperfield is now responsible for the New and Emerging Methods section. Please let James (james.chipperfield@abs.gov.au) know if you would like to contribute to the New and Emerging Methods section in the future. Also, if you have any question which you would like to see answered by an expert, please send it to the new Ask the Experts section editor Ton de Waal from Statistics Netherlands (t.dewaal@cbs.nl). If you are interested in writing a book or software review or suggesting a source to be reviewed, please get in touch with the Book & Software Review section editor Emilio López Escobar from the Quantos Investigación Cuantitativa in Mexico (emilio@quantos.mx). Finally, the country reports should be sent to Peter Wright from Statistics Canada (peter.wright2@canada.ca). If you have any information about conferences, events or just ideas you would like to share with other statisticians – please do go ahead and contact any member of the editorial board of the newsletter.

The Survey Statistician is available for downloading from the IASS website at
http://isi-iass.org/home/services/the-survey-statistician/.

Danutė Krapavickaitė (danute.krapavickaite@vilniustech.lt)

Eric Rancourt (eric.rancourt@canada.ca)

Letter from the President

Dear IASS members!

We are starting 2021 still facing many challenges and adapting to the new circumstances presented to all of us in 2020. Even though, I keep in mind the message of a Brazilian journalist who wrote: “whoever had the idea to cut time into slices, called the year, was a brilliant individual. Industrialized hope, making it work at the limit of exhaustion. Twelve months is enough for any human being to get tired and retreat”. Then “… comes the miracle of renewal and it all starts again, with another number and another desire to believe that from now on everything will be different”. So, being positive about 2021, I would like to highlight IASS successful actions and activities during the last semester of 2020 and compliment the IASS community members who contributed to the good news.

As planned, the IASS webinars, launched in June, became a successful regular activity. Overall, the webinar series was composed of four IASS webinars plus one joint IAOS-IASS webinar, celebrating the World Statistics Day and fostering collaboration between the associations. I thank the organisers, speakers, attendees, and the ISI PO for making it happen. These webinars were part of the ISI and Associations’ Webinar programme and, with financial support from the World Bank (WB) Trust for Capacity Building, all videos are available at https://www.isi-web.org/webinars. The IASS ones can also be found in our website. Besides the webinars, IASS organised a virtual short course in Spanish on Survey Sampling and Survey Methods Topics in R for Latin American attendees. I express my gratitude to Pedro Silva and Guilherme Jacob, course instructors, and Leonardo Trujillo, course organiser. The News and Announcements session contains more information about this course. In addition, Marcel Vieira prepared a short course in English about the same topic that will be available as a self-learning material soon.

For the first time, in August, we held a virtual IASS Annual General Assembly. This was possible since the ISI have been given an exception (in Dutch law) for such meetings to occur virtually in 2020, even if this is not included in the articles of association. In the occasion, an amendment of the IASS Statutes was approved to allow a meeting of the General Assembly of the association to be held at each ISI biennial World Statistics Congress, at another conference or in virtual form. We thank all members who could attend the meeting for their inputs and support. A special thank you goes to Gordon Brackstone (IASS President 2005-2007) who responded to our call to complete information about past council members. Please have a look at (http://isi-iass.org/home/past-committees/).

Last semester was also marked by preparations for the World Statistics Congress 2021 (WSC2021) and the announcement of the 2021 Cochran-Hansen Prize (http://isi-iass.org/home/cochran-hansen-prize/). We ask the members to advertise the award, motivating young statisticians to join the competition. IASS vice-president Isabel Molina is the chair of the C-H Prize Committee. Also, if you have submitted an invited session proposal to the WSC2021, please let Monica Pratesi (IASS representative in the WSC2021 Scientific Programme Committee) and James Chipperfield know about it.

Although 2020 was a difficult year, many people devoted their time to our association. I will not list all of them here, but I would like to acknowledge that for every IASS initiative, committee, webinar, course, its website and each TSS page, there is someone keen to contribute. This is great and it means that IASS matters for all of us. Let us keep this collaborative environment for the sustained development of IASS.

I wish you and your families a year with health, peace, harmony, and that you may be engaged in many gratifying professional projects in 2021.

With my best wishes,

Denise Silva

denisebritz@gmail.com

Report from the Scientific Secretary

It has again been a busy 6 months for me in the role of Scientific Secretary. In this report I introduce the article in the *New and Emerging Methods* section of the Survey Statistician, mention IASS-sponsored webinars that are available for you to watch, mention three invited session proposals for the World Statistics Conference that are sponsored by the IASS, advertise the Cochran-Hansen Prize 2021 and the Waksberg 2022 awards, and list upcoming conferences.

I encourage you to share your views on the functions of the Scientific Secretary with me.

New and Emerging Methods

Professor Li-Chun Zhang gives a thought-provoking account of sampling from a graph, defined as a population of units and connections between them. The concept of connections is applicable to the modern world (e.g., spread of disease and social networks) and it's great that statisticians are working on frameworks for incorporating them in sampling.

The format of *The New and Emerging Methods* articles is 8-10 pages and should cover the presenting challenge, the methods and their application, and the relevance to the development of survey methods. Please contact me if you are interested in writing such an article for future editions of *The Survey Statistician*.

Webinars

At time of writing, the IASS EC has held five webinars during 2020. Three were in response to the Covid-19 pandemic, one on population size estimation, and another to celebrate 2020 World Statistics Day. They are popular, attracting an audience of between 60 and 100. All of these are available on the IASS website (see http://isi-iass.org/home/webinars/). If you would like to contribute to the webinar series or would like to suggest a topic, please contact me.

IASS Website

We are very proud of the IASS website and we encourage you to check it out. It has monthly IASS newsletters, links to IASS webinars and much more! Thanks to Harry Raymond, Australian Bureau of Statistics and IASS Webmaster, for posting the information to our website.

Invited Session Proposals for the WSC 2021 Supported by IASS

If you are an IASS member and have submitted an ISP, please let me and Monica Pratesi (monica.pratesi@unipi.it) know. We are trying to keep track of them. The IASS EC has assisted the UNSD Inter-secretariat Working Group on Household Surveys on their ISP submission called Big Data for Official Statistics - Tips, Tricks and Techniques. I am aware of three ISPs for the WSC 2021 that are supported by the IASS. For your interest, the session's titles and contributors are listed below. Additional information about the sessions will be available in the IASS Newsletter for December 2020.

- *Session: Women in survey sampling research and practice*
  - Organiser: Alina Matei (University of Neuchatel)
  - Chair: Giovanna Ranalli (University of Perugia)
  - Discussant: Monica Pratesi (University of Pisa)
  - Speakers: Frauke Kreuter (University of Maryland), Claudia Rivera Rodriguez (University of Auckland, New Zealand), and Anne Ruiz-Gazen (University of Toulouse 1)

- *Session: Counting People with Administrative Data instead of a Traditional Census*
    - Chair: Prof. James Brown, (University Technology Sydney)
    - Discussant: Prof. Li-Chun Zhang (Southampton University)
    - Speakers: Ms Alison Whitworth (Office of National Statistics), Ms. Nicoletta Cibella (National Institute of Statistics), Ms. Antonella Bernardini (National Institute of Statistics), Dr. Ahmad Hleihel (Israel Central Bureau of Statistics)
- *Session: The Evolution of the 21st Century Census of Population and Housing*
    - Chair: Dr. James Chipperfield (Australian Bureau of Statistics)
    - Discussant: Prof. James Brown (University Technology Sydney)
    - Speakers: Ms. Gemma Van Halderen, (United Nations), Ms. Ms Valérie Roux (INSEE), Mr. Gary Dunnet (Statistics New Zealand)

Invitation to apply for Prizes and Awards

Cochran-Hansen Prize 2021: In celebration of its 25th anniversary in 1999, the International Association of Survey Statisticians (IASS) established the Cochran-Hansen Prize, which is awarded every two years for the best paper on survey research methods submitted by a young statistician from a developing or transition country. In the 2021 edition, the Cochran-Hansen Prize consists of research expenses in a total of €1500 to be executed until 15 February 2023. For details see http://isi-iass.org/home/cochran-hansen-prize/cochran-hansen-prize-2021/.

2022 Waksberg Award: The journal Survey Methodology has established an annual invited paper series in honour of the late Joseph Waksberg to recognize his outstanding contributions to survey statistics and methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey statistics and methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work. For more details see http://isi-iass.org/home/waksberg-award/.

WB capacity Building Fund for Analysis of survey data using R short course

The IASS was funded by the World Bank to organise the delivery of courses on Analysis of survey data using R during 2020 to developing countries. We are pleased to announce that the IASS has arranged two such courses. Marcel Vieira from Brazil has prepared a short course (video and slides) in English that can be used as a self-learning activity and will be available on the ISI webinar. Prof. Pedro Luis do Nascimento Silva, with support from teaching assistant Guilherme Anthony Pinheiro Jacob, also delivered a course in Spanish.

Conferences

If you would like to advertise a conference, workshop, or course that would be of interest to IASS members please let me know.

James Chipperfield

James.chipperfield@abs.gov.au

IASS Scientific Secretary

News and Announcements

ISI World Statistics Congress 2021 goes VIRTUAL!

The International Statistical Institute's biennial 'World Statistics Congress 2021 - The Hague' (https://isi2021.org/), will be held virtually in July 2021.

This is an opportunity for the ISI to host a more inclusive conference than ever before. It will allow us to have the greatest ever participation, reaching out to a wide diversity of members who would not otherwise be able to afford to attend an international conference, at lower costs, and with a lower carbon footprint. We hope to see a large number of first-time participants, including many early career statisticians and students along with colleagues from developing countries.

Holding a WSC in 2021 will maintain the continuity of WSCs. There will still be a strong link with The Hague, the original venue, in the virtual conference. The ISI will take the opportunity to develop new formats and harness new technology to modernise our meeting.

Please visit the website for information on the scientific programme (https://isi2021.org/scientific-programme.html), including the timetable, and for registration (https://isi2021.org/registration.html) categories and fees.

ISI World Statistics Congress 2023: OTTAWA is chosen!

The ISI is (also) pleased to announce that the World Statistics Congress 2023 will be held in Ottawa, Canada. The Congress will be held from 15 to 20 July, in the Shaw Centre (https://www.shaw-centre.com/), in the heart of the city.

The venue was chosen after competitive bids from a number of venues in Canada were assessed. We are grateful to the Statistical Society of Canada and to Statistics Canada for their support. We are confident that the Congress will build on the experience of WSC 2021 and will include a virtual element.

We look forward to meeting up with colleagues and friends again in Canada!

Announcement: New editor of Survey Methodology

Statistics Canada is pleased to announce the appointment of Jean-François Beaumont, senior statistical advisor at Statistics Canada, as the new Editor of *Survey Methodology*, effective January 1st, 2021. Jean-François Beaumont has been associated with the Journal for 20 years; he first served as assistant editor from 2000-2010, and then as associate editor from 2010-2020. He published several papers in a number of peer-reviewed journals, including nine in *Survey Methodology*.

Jean-François Beaumont succeeds Dr. Wesley Yung, who has been the Editor of *Survey Methodology* since 2015. Dr. Yung will remain a member of the Management Board and will continue to contribute to the Journal as associate editor. We deeply thank him for his valuable contributions as editor over the past five years. Amongst other achievements, he steered the journal into publishing more than 80 papers. He also led the journal in publishing a special issue as part of a collaboration with the *International Statistical Review* in honour of Prof. J.N.K. Rao's contributions. Dr. Yung was also instrumental in the implementation of the ScholarOne online manuscripts submission system in 2019.

Survey Methodology has evolved tremendously over the past 45 years under the leadership of the previous editors, starting with Dr. M.P. Singh (1975-2005), and then John Kovar (2006-2009), Dr. Michel Hidiroglou (2010-2015) and recently Dr. Wesley Yung (2016-2020). In the first years after its creation, the Journal published papers mainly from Statistics Canada's authors, but it soon acquired the worldwide reputation of a high-quality journal for survey statisticians and methodologists under the dedicated editorship of Dr. M.P. Singh. *Survey Methodology* continued to thrive under the subsequent editors to become the journal we know today. It nowadays publishes innovative theoretical or applied statistical research papers from international authors on issues relevant to the activities of National Statistical Offices.

While the contributions of previous editors were invaluable in making *Survey Methodology* a journal with international stature, we look forward to see where Jean-François Beaumont will take the Journal in the years to come. There is a new publication landscape with many players in the field and authors have heighten expectations in terms of review expediencies. While navigating this new ecosystem, we are confident that *Survey Methodology* will continue to flourish under the editorship of Jean-François Beaumont and that it will keep its same essential core value of scientific rigour.

Survey Methodology is available for free at www.statcan.gc.ca/SurveyMethodology.

Analysis of survey data using R short course

The short course on 'Analysis of survey data using R' (**Análisis de datos de encuestas usando R**) was held in six remote sessions of 3 hours each, on November 13, 14, 20, 21, 27 and 28. The course was taught by Prof. Pedro Luis do Nascimento Silva, with support from teaching assistant Guilherme Anthony Pinheiro Jacob. Teaching was carried out in Spanish, with course notes and all other relevant material (R code, data sets, references etc.) provided to the participants.

A total of 47 participants completed the course, with 1 from Panama, 4 from Peru and 42 from Colombia.

The course was hosted by the Universidad Nacional de Colombia, sede Bogotá, under the leadership of Prof. Leonardo Trujillo Oyola. It had support from several scientific and professional organizations, namely:

- International Statistical Institute (ISI) – Statistical Capacity Building Committee;
- International Association of Survey Statisticians (IASS);
- Instituto InterAmericano de Estadística (IASI);
- Sociedad Colombiana de Estadística;
- Escola Nacional de Ciências Estatísticas (ENCE/IBGE).

The course covered aimed to develop participants' capacity to analyse complex sample survey data using modern methods and tools available in the R statistical software, including tools for weighting survey data, computing point and standard error estimates, model fitting, selection and diagnostics. Numerous examples using real survey data were provided and formed an integral part of the training.

Pedro Luis do Nascimento Silva

December 2020

A tribute to Ken Brewer

Ken Brewer, one of the giants of survey sampling, passed away peacefully in the early hours of 3 January, 2021 in Canberra.

Ken was a great scholar, very wise and showed interests in many areas of statistical applications, on which he published extensively. He had also published two books and over 100 papers on survey sampling, many of which were highly cited.

Apart from his academic achievements, Ken was also a very kind man. He never stopped sharing his insights and knowledge on statistics with the more junior colleagues. He also generously offered help to many others in the different stages of their career - I was one of those who benefitted from his advice to settle in Canberra in the mid 1980's as a new immigrant and, since that time, learned a lot from working alongside with him.

Summarising one of Ken's most noticeable contributions very well when, announcing Ken's passing away to the subscribers of the Survey Research Methodology Section of the American Statistician, Phil Knot said:

"One of Ken Brewer's most remarkable achievements was a paper effectively summarizing over 20 years of the design-based vs model-based debate, which ranged from Richard Royall's infamous 1970 Biometrika paper till the model-assisted synthesis coined and described in Särndal, Swensson, and Wretman's 1992 Book. Ken published this paper in the Journal of Australian Statistics in 1963."

In many areas of research, Ken was ahead of his time.

Ken was an excellent researcher, colleague, friend and teacher to many of us. Ken will be sorely missed.

Siu-Ming Tam

Honorary Professorial Fellow, University of Wollongong

The Survey Statistician, 2021, Vol. 83, 11.-22.

Ask the Experts

Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies?

Jean-Francois Beaumont[1] **and J. N. K. Rao**[2]

[1] Statistics Canada, jean-francois.beaumont@canada.ca

[2] Carleton University, jrao@math.carleton.ca

Abstract

There is a growing interest in National Statistical Offices to produce Official Statistics using non-probability sample data, such as big data or data from a volunteer web survey, either alone or in combination with probability sample data. The main motivation for using non-probability samples is their low cost and respondent burden, and quick turnaround since they allow for producing estimates shortly after the information needs have been identified. However, non-probability samples are not a panacea. They are well known to produce estimates that may be fraught with significant selection bias. We first discuss this important limitation, along with an illustration, and then describe some remedies through inverse probability weighting or mass imputation. We also discuss how to integrate data from probability and non-probability samples through the Fay-Herriot model used in Small Area Estimation. We conclude with a few remarks on some challenges that statisticians are facing when implementing data integration methods.

Key words: Big data, Inverse probability weighting, Mass imputation, Selection bias, Small area estimation.

1 Introduction

Large-scale sample surveys, based on properly designed probability samples, have long been used to obtain reliable estimates of population totals, means and other descriptive parameters. Repeated sampling (or design-based approach) is widely employed for this purpose ever since the landmark paper by Neyman (1934) laid the theoretical foundations of the design-based approach to inference. Attractive features of this approach include design-consistent estimation of the parameters and associated mean squared errors, and normal theory confidence intervals on the parameters, that are valid, at least for large enough samples, “whatever the unknown properties of the finite population” (Neyman, 1934).

Efficient sampling designs minimizing cost subject to specified precision of estimators and “optimal” estimation of parameters taking advantage of supplementary information, such as censuses and administrative data, were proposed. Methods for combining two or more independent probability samples were also developed to increase the efficiency of estimators for a given cost. Many important large-scale surveys conducted by official statistical agencies, such as labor force, business, and agricultural surveys, continue to follow the traditional approach.

In the early days of probability sampling evolution, surveys were generally much simpler than they are nowadays, and data were largely collected through personal interviews or through mail questionnaires followed by personal interviews of nonrespondents. Physical measurements were also used. Data collection issues received much attention in recent years to control costs and maintain response rates using new modes of data collection adapted to technological changes. Despite those efforts to collect designed data under probability sampling, response rates are decreasing, and costs and response burden are increasing. On the other hand, largely due to technological innovations, large amounts of inexpensive data, called big data or organic data, and data from nonprobability samples (especially self-selection web surveys) are now accessible. Big data include transaction data, social media data, scrape data from websites, sensor data and satellite images. Such data have the potential of providing estimates in near real time, unlike traditional designed data collected from probability samples.

Statistical agencies publishing official statistics are now undertaking modernization initiatives by finding new ways to integrate data from a variety of sources and produce “reliable” official statistics quickly. However, naïve use of data from nonprobability samples or big data can lead to serious selection bias problems. Without using suitable adjustments to account for selection bias, it can lead to the big data paradox: the bigger the data, the surer we fool ourselves, as demonstrated in Section 3 (Meng, 2018) and in the illustration in Section 4. We discuss some remedies in Section 5 to reduce the pitfalls arising from making inferences from nonprobability samples or big data. We show that the methods designed for making inferences from probability samples can be adapted for nonprobability samples. In Section 6, we discuss how small area estimation techniques can be used to integrate data from probability and non-probability samples. We provide a few concluding remarks in the last section on some of the practical challenges that require further thinking.

2 Probability sampling

2.1 Design-based approach

A distinctive feature of a probability sample A is that it ensures every unit i in the finite population U has a known nonzero inclusion probability π_i, leading to design weights $d_i = \pi_i^{-1}$ and a basic design-unbiased expansion estimator $\hat{Y} = \sum_{i \in A} d_i y_i$ of the finite population total $Y = \sum_{i \in U} y_i$ of a variable of interest y. Extensive research was conducted to improve the efficiency of the expansion estimator through the use of auxiliary variables $\mathbf{x}$ with known population totals $\mathbf{X}$. This is accomplished at the design stage through probability proportional to size sampling and stratification or at the estimation stage through ratio or regression estimation or both. The resulting improved estimators are not necessarily design-unbiased, but they are design-consistent in large samples. A well-known example is the ratio estimator $\hat{Y}_r = (\hat{Y} / \hat{X})X$ extensively used in practice, where $\hat{X}$ is the expansion estimator of the known total X. The ratio estimator may be expressed as a calibration estimator $\hat{Y}_r = \sum_{i \in A} w_i y_i$ with calibration weights $w_i = (X / \hat{X})d_i$ that ensure the calibration property $\sum_{i \in A} w_i x_i = X$. This property ensures that the ratio estimator agrees with the known total X when y_i is replaced by x_i.

Extensive research has been undertaken to extend calibration estimation to a vector of auxiliary variables $\mathbf{x}$ with known totals $\mathbf{X}$. A simple way of constructing calibration weights is to minimize a chi-squared distance measure between d_i and w_i for $i \in A$ with respect to w_i subject to calibration constraints $\sum_{i \in A} w_i \mathbf{x}_i = \mathbf{X}$. Post-stratification is an important special case which occurs when all elements of $\mathbf{x}_i$ but one is equal to 0. Calibration estimation has attracted the attention of users due to its model-free property and its ability to produce a common set of weights not depending on the study variable. Van den Brakel and Bethlehem (2008) note that the calibration weights are “very attractive to produce timely official statistics in a regular production environment”. The calibration approach has the potential to adjust for selection bias of non-probability samples, as noted in Section 5.

Design-consistent variance estimation under probability sampling applicable to general descriptive parameters, leading to normal theory confidence intervals on the parameters, also received a lot of attention. Methods proposed include Taylor linearization and replication methods, such as the jackknife and the bootstrap, taking account of the design features.

2.2 Unit nonresponse

Bias due to unit nonresponse in a probability sample received considerable attention, and promising remedies were proposed under a random response model. Under this model, the units in the population are assumed to respond independently if selected in the sample with unknown probabilities q_i, $i \in U$. Suppose we select a simple random sample (SRS) of size n and use the sample mean of respondent values as the estimator of the population mean $\bar{Y}$. Then, under the above design-model set up, the bias of the naïve estimator is approximately equal to $B_q =$ $(R_{qy} S_q S_y) / \bar{Q}$, where R_{qy} is the finite population correlation between the study variable and the response probability, S_q and S_y denote the standard deviations of the response probabilities q_i and values y_i of the study variable, and $\bar{Q}$ is the population mean of the response probabilities (Bethlehem, 2020). It follows from the above bias expression that the bias increases with R_{qy}, S_q and decreasing response rate, $\bar{Q}$. The bias disappears if the response probabilities are identical ($S_q = 0$) or nonresponse is not selective ($R_{qy} = 0$). The latter case, called missing completely at random (MCAR), seldom holds in practice.

Success of any adjustment for nonresponse bias depends on the availability of auxiliary variables $\mathbf{z}$ for all the sampled units that are closely related to the study variable y. In, for instance, Scandinavian countries, population registers are often used to extract $\mathbf{z}$ for all the units in the sample. Missing at random (MAR) assumption plays a dominant role in estimating the response probabilities (propensities) q_i. Under MAR, $q_i = q(\mathbf{z}_i, \boldsymbol{\alpha})$ for a specified function $q(.)$ depending only on the observed $\mathbf{z}_i$ and a parameter $\boldsymbol{\alpha}$. A bias-adjusted expansion estimator of the total Y under MAR is constructed as $\hat{Y}_q = \sum_{i \in A(r)} (d_i / \hat{q}_i) y_i$, where $A(r)$ is the sample of respondents and $\hat{q}_i = q(\mathbf{z}_i, \hat{\boldsymbol{\alpha}})$ is the estimated response probability. The parameter $\boldsymbol{\alpha}$ is estimated from the sample data $\{(\delta_i, \mathbf{z}_i), i \in A\}$, where δ_i is the response indicator. Logistic regression models, $\log\{q_i / (1 - q_i)\} = \mathbf{z}_i' \boldsymbol{\alpha}$, are commonly used for this purpose. The bias-adjusted estimator $\hat{Y}_q$ is design-model consistent, provided the response propensity model is correctly specified. The estimator $\hat{Y}_q$ becomes unstable when some of the estimated response propensities become small. To avoid this problem, weighting classes can be formed, based on quantiles of the estimated response probabilities, and then use a post-stratified estimator by assuming uniform response probabilities within classes.

3 Non-probability samples: selection bias

Consider a non-probability sample B of known size N_B with data $\{(i,y_i), i \in B\}$ and a participation indicator δ_i taking the value 1 if the population unit i belongs to B and 0 otherwise. In the absence of auxiliary information $\mathbf{z}$, the estimator of $\bar{Y}$ is taken as the sample mean $\bar{y}_B = N_B^{-1}\sum_{i \in U} \delta_i y_i$. T0he estimation error $\bar{y}_B - \bar{Y}$ may be expressed as the product of three terms: (1) $corr(\delta, y) = R_{\delta,y}$, called data quality, (2) square root of $(1-f_B)/f_B$ with $f_B = N_B / N$, called data quantity and (3) square root of the population variance S_y^2, called problem difficulty (Meng, 2018). The data quality term plays the key role in determining the estimation error and it is approximately zero on the average under SRS. Note that we have not used a random participation mechanism that assumes the participation indicators δ_i are random and independent with non-zero participation probabilities $q_i = P(\delta_i = 1)$; for simplicity, we use the same notation for response propensities and participation probabilities. Under the random participation model, the bias of $\bar{y}_B$ given by $E_\delta(\bar{y}_B - \bar{Y})$ has the same expression as in the case of nonresponse. The subscript δ indicates that the expectation is taken with respect to the participation model. However, Bethlehem (2020) notes that in practical situations the two values can be substantially different. He gives an example from Netherlands where $\bar{Q}$ is around 60% for probability surveys compared to 1.5% for self-selection web surveys, even though 170,000 people completed the questionnaire in the web survey. This example suggest that self-selection surveys can suffer from the risk of a much larger bias.

The model mean squared error $MSE_\delta(\bar{y}_B)$ conditional on the sample size N_B may be expressed as the product of three terms: (1) $E_\delta(R_{\delta,y}^2)$, called the data defect index, (2) drop-out odds $(1-f_B)/f_B$ and (3) degree of uncertainty S_y^2. Note that MSE is affected by the sampling fraction f_B and not the sample size N_B. As a result, a relatively small simple random sample of size n can achieve the same MSE. For example, suppose N_B is five million and N is ten million leading to $f_B = 1/2$, and the average correlation $E_\delta(R_{\delta,y})$ is as small as 0.05. Then the “effective” sample size of the big data is less than 400. Moreover, the confidence interval, treating the big data as a simple random sample, has a small chance of covering the true mean $\bar{Y}$ because it is centered at a wrong value due to the induced bias. We know this phenomenon under probability sampling when the ratio of bias to standard error is large. For a design-consistent estimator, such as a ratio estimator, the bias ratio goes to zero as the sample size increases.

For simplicity, we assumed the absence of measurement errors in the nonprobability sample B. This is often not the case with found data from online sources, such as Facebook, where people may actively lie, and the expected bias due to measurement errors could be large. Biemer (2019) extended Meng’s model to show that the bias due to measurement errors could significantly inflate the total MSE.

Unlike in the case of nonresponse in a probability sample A, auxiliary variables $\mathbf{z}$ attached to the units not participating in the nonprobability sample B are seldom available. As a result, it is not possible to estimate participation probabilities from sample B alone and make bias adjustments. In Section 5 we study some methods of estimating participation probabilities by supplementing the data $\{(i,\mathbf{z}_i), i \in B\}$ with the data $\{(i,\mathbf{z}_i), i \in A\}$ obtained from an independent probability sample A observing $\mathbf{z}$ and possibly different study variables. This set up has received a lot of attention in the recent literature, but we focus on a pseudo-likelihood method proposed by Chen, Li and Wu (2019) and a mass imputation method of Rivers (2007).

4 An illustration using real data

After the beginning of the COVID-19 lockdown in March 2020, Statistics Canada conducted a series of crowdsourcing experiments to respond to urgent information needs about the life of the Canadian population. A crowdsourcing sample can be defined as any non-probability sample of volunteers, who typically provide information through an online application. Statistics Canada's crowdsourcing data were collected by posting questionnaires on its website on different topics at regular intervals. The main advantages of crowdsourcing are its low cost and quick turnaround since estimates can be released within a couple of weeks after the information needs have been determined. This timeliness was deeply needed in a pandemic time. The first crowdsourcing experiment was viewed as a success considering that around 240,000 persons participated. However, the number of participants in the subsequent crowdsourcing experiments was smaller, but often reached over 30,000 participants. As pointed out in Section 3, ignoring possible measurement errors, the main drawback of non-probability surveys of volunteers is the selection bias, also called participation bias. To account for this bias, it was decided to apply post-stratification weighting with post-strata defined by the cross-classification of province, age group and sex. Renaud and Beaumont (2020) provide greater detail on crowdsourcing experiments conducted by Statistics Canada.

In parallel, Statistics Canada also started a shorter series of probability web panel surveys: the Canadian Perspective Survey Series (CPSS). The CPSS sample is obtained from past rotation groups of the Labour Force Survey (LFS), which is the most important social survey conducted by Statistics Canada except for the Census. The CPSS initial probability sample is relatively large with over 30 000 selected persons but the overall recruitment/response rate is usually quite low at around 15%, and the resulting number of respondents is just slightly over 4,000. Greater detail on the CPSS can be found in Baribeau (2020).

In June 2020, some participants from previous crowdsourcing experiments were randomly chosen and sent the same questionnaire as CPSS respondents. This allowed for a comparison of estimates from both the CPSS probability sample and this crowdsourcing non-probability sample. We provide some results for the variable *education* as this variable is also available in the LFS, which has generally a response rate around 80%, and is treated as our gold standard. The June 2020 CPSS contained 4,209 respondents whereas the corresponding crowdsourcing sample had 31,505 participants, and the LFS had 87,970 respondents. We computed LFS, CPSS and crowdsourcing estimates of proportions in different education categories (see Table 1 for the description of categories) for Canada (see Figure 1) and for a province of Canada (see Figure 2). Normal 95% confidence intervals were also computed for the LFS and CPSS. We considered two versions of crowdsourcing estimates: unadjusted and post-stratified. The unadjusted crowdsourcing estimates were obtained using an estimation weight equal to 1 for every participant, and the post-stratified crowdsourcing estimates were obtained by using a post-stratification weight with post-strata defined by the cross-classification of province, age group and sex. Three main conclusions can be drawn from Figure 1:

i) The crowdsourcing sample seems to significantly over-represent those with a university degree.
ii) Post-stratification by province, age group and sex have little impact on the estimates, and thereby on the selection bias.
iii) The CPSS estimates are closer to LFS estimates although with larger confidence intervals due to the smaller sample size.

The same conclusions can be drawn with the provincial estimates in Figure 2, the main difference being that confidence intervals are wider. For the province considered, the number of respondents in the LFS and CPSS are 4,734 and 231, respectively, and the number of crowdsourcing participants is 1,716.

Table 1: *Categories of education*

Education categories	Description
0	Grade 8 or lower
1	Grade 9 - 10
2	Grade 11 - 13, non graduate
3	Grade 11 - 13, graduate
4	Some post-secondary education
5	Trades certificate or diploma
6	Community college, CEGEP, etc.
7	University certificate below Bachelor's
8	Bachelor's degree
9	Above Bachelor’s degree

A caveat must be mentioned about conclusion (iii): nonresponse weighting in the CPSS used the study variable, education, as one of the auxiliary variables. This may explain the small difference between the CPSS and LFS estimates for that variable, especially at Canada level. Ideally, nonresponse weighting would be done again by omitting the auxiliary variable education. This would better allow us to appreciate the accuracy of the CPSS probability sample. Unfortunately, this could not be done before the publication of this paper.

Conclusion (ii) indicates that province, age and sex are insufficient for significantly reducing the selection bias. More powerful auxiliary variables are needed for this purpose with possibly more sophisticated methods such as propensity score weighting (Chen, Li and Wu, 2019) or sample matching (Rivers, 2007) discussed in Section 5. In practice, the variable education itself can be used to reduce the selection bias for other study variables. Preliminary results suggest that it is a powerful auxiliary variable although insufficient for entirely removing the selection bias.

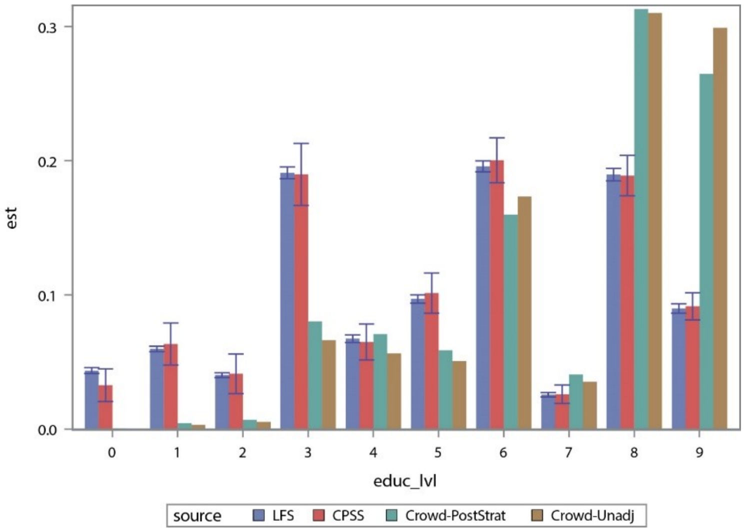


Figure 1: *Estimates of proportions in different education categories for Canada from different sources: a large probability survey (LFS), a small probability survey (CPSS), a non-probability survey (Crowd-Unadj) and a non-probability survey with post-stratification weighting by province, age group and sex (Crowd-PostStrat)*

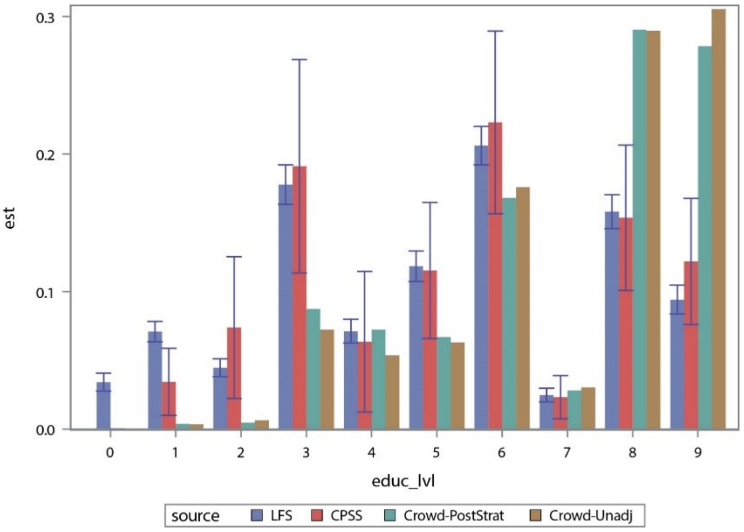


Figure 2: *Estimates of proportions in different education categories for a Canadian province from different sources: a large probability survey (LFS), a small probability survey (CPSS), a non-probability survey (Crowd-Unadj) and a non-probability survey with post-stratification weighting by province, age group and sex (Crowd-PostStrat)*

5 Methods for adjusting selection bias

Our data for estimating participation probabilities consist of the common auxiliary variables $\mathbf{z}$ observed in both samples A and B. We assume a model for the participation probabilities $q_i = q(\mathbf{z}_i, \boldsymbol{\alpha})$ for specified $q(.)$ under the MAR assumption. Chen, Li and Wu (2019) estimate the population log-likelihood $l(\boldsymbol{\alpha})$ based on the sample B by using the data $\{(i, \mathbf{z}_i), i \in A\}$ and associated design weights $d_i, i \in A$, leading to a pseudo-likelihood $\hat{l}(\boldsymbol{\alpha})$. For the commonly used logistic regression model $\log\{q_i / (1 - q_i)\} = \mathbf{z}_i'\boldsymbol{\alpha}$, the corresponding pseudo-score equation reduces to

$$\hat{\mathbf{U}}(\boldsymbol{\alpha}) = \sum\nolimits_{i \in B} \mathbf{z}_i - \sum\nolimits_{i \in A} d_i q(\mathbf{z}_i, \boldsymbol{\alpha}) \mathbf{z}_i = \mathbf{0}. \qquad (1)$$

Equation (1) is solved using the Newton-Raphson iteration procedure with a starting value $\boldsymbol{\alpha}^{(0)} = 0$, leading to an estimator $\hat{\boldsymbol{\alpha}}$ and corresponding estimated probability $\hat{q}_i = q(\mathbf{z}_i, \hat{\boldsymbol{\alpha}})$. The resulting inverse probability weighted (IPW) estimator of the mean is $\bar{y}_{B,q} = \sum\nolimits_{i \in B} \omega_i y_i / \sum\nolimits_{i \in B} \omega_i$, where $\omega_i = \hat{q}_i^{-1}$. The corresponding ratio estimator of the total is $\hat{Y}_{B,q} = N \bar{y}_{B,q}$, but the population size N may not be known in practice unless the probability sample is drawn from a list frame. Conventional calibration estimation can also be used to estimate $\boldsymbol{\alpha}$ by solving

$$\sum\nolimits_{i \in B} [q(\mathbf{z}_i, \boldsymbol{\alpha})]^{-1} \mathbf{z}_i = \sum\nolimits_{i \in A} d_i \mathbf{z}_i. \qquad (2)$$

A drawback of the IPW estimator is that it is sensitive to misspecification of the participation probability model, especially when some of the $\hat{q}_i, i \in A,$ are small. One way to achieve some robustness is to use a double robust (DR) estimator that requires modeling each study variable y. Suppose we assume a working population model, $E_m(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta}), i \in U$, where E_m denotes model expectation. The choice of auxiliary variables $\mathbf{x}_i$ may include the common variables $\mathbf{z}_i$ and other available auxiliary variables in the sample B that are predictive of the study variable y_i. We further assume the population model holds for the sample B. Using the data from the sample B we obtain an estimator $\hat{\boldsymbol{\beta}}$ which is consistent for $\boldsymbol{\beta}$ under the assumed model. Then a DR estimator of the total is given by

$$\hat{Y}_{DR.q} = \sum\nolimits_{i \in B} \omega_i (y_i - \hat{m}_i) + \sum\nolimits_{i \in A} d_i \hat{m}_i \ , \tag{3}$$

where $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ denote the predicted values under the model. The estimator (3) is DR in the sense it is consistent if either the model for the participation probabilities or the model for the study variable is correctly specified. The DR estimator of the mean is given by $\bar{y}_{DR,q} = \hat{Y}_{DR,q} / \hat{N}_q$, where $\hat{N}_q = \sum\nolimits_{i \in B} \omega_i$ is the IPW estimator of the population size N. DR estimators have been used in the context of nonresponse in a probability sample (Kim and Haziza, 2014).

Chen, Li and Wu (2019) provided asymptotically valid variance estimators of the DR estimator and the IPW estimator under the assumed models. They also conducted a limited simulation study on IPW, DR and some alternative estimators of the mean. As expected, the DR estimator performed well in terms of relative bias (RB) and MSE when either the assumed model for participation probabilities or the model on the study variable is correctly specified. On the other hand, the IPW estimator performs poorly when the model on the participation probabilities is incorrectly specified. Chen, Li and Wu (2019) did not study the case when both models are incorrectly specified. It would be interesting to develop multiple robust (MR) estimators, along the lines of Chen and Haziza (2017) in the context of unit nonresponse, by specifying multiple models on the study variable and multiple models for the participation probability. Under this set up, an estimator is MR if it performs well when at least one of the candidate models is correctly specified. Simulation results in the context of nonresponse indicated that MR estimators tend to perform well even when the candidate models are all incorrectly specified. It would be interesting to study MR estimation in the context of nonprobability samples.

A popular method is based on non-parametric mass imputation that avoids the specification of the mean function $E_m(y_i \mid \mathbf{z}_i)$ based on common auxiliary variables $\mathbf{z}_i$ (Rivers, 2007). In this method, for each unit $i \in A$, a nearest neighbor (NN) to the associated $\mathbf{z}_i$ is found from the donor set $\{(i, \mathbf{z}_i), i \in B\}$, say $(l, \mathbf{z}_l)$, and the corresponding y_l is used as the imputed value $y_i^*(= y_l)$. Euclidean distance is commonly used to find the NN. The mass imputed estimator of the total is then given by

$$\hat{Y}_{d,I} = \sum\nolimits_{i \in A} d_i y_i^* . \tag{4}$$

The estimator (4) is based on real donor values. However, it is not exactly model-design unbiased unless $E_m(y_i^* \mid \mathbf{z}_i) = E_m(y_i \mid \mathbf{z}_i)$ for $i \in A$. The NN estimator of the mean is given by $\bar{y}_{d,I} = \hat{Y}_{d,I} / \hat{N}_d$, where $\hat{N}_d = \sum\nolimits_{i \in A} d_i$ is the design-unbiased estimator of the total. Under certain regularity conditions, including smoothness of the mean function and MAR assumption, Yang and Kim (2020) showed that the NN estimator (4) behaves asymptotically similar to the design-unbiased estimator with the imputed values in (4) replaced by the unobserved true values y_i of the units $i \in A$.

An advantage of the estimator (4) is that it can be readily implemented when non-probability samples are observed frequently over time leading to time-varying values of the study variable, provided the common auxiliary variables do not vary over time. We simply find the NN for the units in the probability sample from the current donor set and calculate (4).

6 Small Area Estimation

Data integration methods in Section 5 can be applied in the scenario where the study variable y is observed in the non-probability sample B, but completely missing in the probability sample A. The reduction of the selection bias relies on the availability of a vector of auxiliary variables $\mathbf{z}$ observed in both samples. In this section, we study a different scenario. We consider the case where the study variable is observed in the probability sample A, and the non-probability sample contains a vector of auxiliary variables associated with y, such as a proxy for y, or perhaps even y itself. Design-unbiased estimators of finite population parameters can be obtained by ignoring all non-probability sample data. Therefore, the selection bias is not an issue in this context. The objective is to take advantage of the information contained in the non-probability sample to improve the efficiency of design-unbiased estimators through model assumptions that link data from both samples. This is the scope of Small Area Estimation (SAE) techniques (Rao and Molina, 2015). We focus below on the area level model of Fay and Herriot (1979). Unit level models are also used in SAE but require the auxiliary variables to be observed in the probability sample, possibly through record linkage, and this may not be feasible.

Suppose that we are interested in estimating J population parameters, θ_j, $j=1,...,J$, where the subscript j refers to J disjoint population domains of interest. For instance, θ_j could be the unemployment rate in a certain area j. Let us denote by $\hat{\theta}_j$, a design-unbiased estimator of θ_j obtained from the probability sample A. From this design-unbiasedness property, we can write the so-called sampling model as

$$\hat{\theta}_j = \theta_j + e_j ,$$

where e_j is the sampling error such that $E_p\left(e_j\right)=0$, $\text{var}_p\left(e_j\right)=\psi_i$ and ψ_i is the design variance of $\hat{\theta}_j$. The subscript p indicates that the expectation and variance are taken with respect to the probability sampling design. In practice, $\hat{\theta}_j$ is rarely exactly design-unbiased but is assumed to be at least approximately design-unbiased. The sampling errors are also assumed to be normally distributed and mutually independent even when the design strata do not coincide with the domains.

The above sampling model is complemented with a linking model that relates the population parameter θ_j to a vector of auxiliary variables $\mathbf{z}_{B,j}$. This vector of auxiliary variables may come from many different sources; in practice, they are often administrative sources. Here, we focus on the use of a non-probability sample as the source of auxiliary information. For instance, a proxy for θ_j could be computed from the non-probability sample B, if a variable similar to y is observed, and used as one component of $\mathbf{z}_{B,j}$. The following linking model is often used:

$$\theta_j = \mathbf{z}'_{B,j}\boldsymbol{\beta} + b_j v_j ,$$

where b_j are constant to account for possible heteroscedasticity, v_j are mutually independent errors that follow the normal distribution with $E_m\left(v_j\right)=0$ and $\text{var}_m\left(v_j\right)=\sigma_v^2$, and $\boldsymbol{\beta}$ and σ_v^2 are unknown model parameters. The subscript m indicates that the expectation and variance are taken with respect to the model, treating the vectors $\mathbf{z}_{B,j}$ as fixed.

The Fay-Herriot model is obtained by combining the sampling and linking model as

$$\hat{\theta}_j = \mathbf{z}'_{B,j}\boldsymbol{\beta} + a_j ,$$

where $a_j = b_j v_j + e_j$ .It is straightforward to show that $E_{mp}\left(a_j\right)=0$ and $\text{var}_{mp}\left(a_j\right)=b_j^2\sigma_v^2+\tilde{\psi}_j$, where $\tilde{\psi}_j = E_m\left(\psi_j\right)$ is called the smooth design variance of $\hat{\theta}_j$ (Hidiroglou, Beaumont and Yung, 2019). The Empirical Best (EB) predictor of θ_j under the Fay-Herriot model is

$$\hat{\theta}_j^{EB} = \hat{\gamma}_j \hat{\theta}_j + \left(1 - \hat{\gamma}_j\right) \mathbf{z}'_{B,j} \hat{\boldsymbol{\beta}},$$

where $\hat{\gamma}_j = b_j^2 \hat{\sigma}_v^2 \Big/ \left(b_j^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_j\right)$, and $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_v^2$ and $\hat{\tilde{\psi}}_j$ are estimators of $\boldsymbol{\beta}$, σ_v^2 and $\tilde{\psi}_j$, respectively. The vector $\boldsymbol{\beta}$ is typically estimated by generalized least squares and a number of methods exist for the estimation of σ_v^2 (see Rao and Molina, 2015), the most common being restricted maximum likelihood. Hidiroglou, Beaumont and Yung (2019) use a log-linear smoothing model and a method of moments for the estimation of $\tilde{\psi}_j$.

Chatrchi, Gauvin and Ding (2020) recently applied the Fay-Herriot model to the estimation of trip spending by visitors to Canada for 242 domains defined by the cross-classification of region and the country of origin of visitors. The direct estimator $\hat{\theta}_j$ is obtained from the Visitor Travel Survey (VTS) and a proxy for θ_j is obtained from payment processors' data. These data are subject to coverage issues and are conceptually different from the y variable measured in the VTS. Yet, they were able to observe significant precision improvements by using the Fay-Herriot model, especially for the smallest domains. Their initial analyses showed the evidence of a nonlinear relationship between $\hat{\theta}_j$ and the proxy for θ_j. A piecewise linear model appears to have addressed this issue satisfactorily. Rao (2020) describes other SAE applications, where authors have used non-probability data and big data as a source of auxiliary information in a SAE model.

7 Concluding remarks

National Statistical Offices have been increasingly considering big data and volunteer web surveys in recent years as an alternative to conducting probability surveys. The main advantages of these non-probability samples over probability samples are their low cost and respondent burden, as well as their quick turnaround, i.e., the gap between the determination of information needs and the release of estimates is quite small for non-probability samples.

It is well known that the use of a non-probability sample alone may lead to biased estimates of finite population parameters due to measurement errors and selection bias (e.g., Rao, 2020; Beaumont, 2020). As a result, it does not seem advisable to use non-probability data to produce official statistics without complementing them with probability survey data. In Section 5, we discussed data integration methods that reduce the bias by combining non-probability and probability survey data. Although these methods are useful to achieve bias reductions, they rely strongly on the MAR assumption. Practically speaking, this assumption implies that the participation indicator is independent of the study variable y after conditioning on $\mathbf{z}$. A powerful set of auxiliary variables, associated with both the participation indicator and the study variable, is key to make this assumption plausible. Before probability and non-probability surveys are conducted, it is thus useful to give some thoughts on the inclusion of proper variables to be collected in both surveys for the purpose of making the MAR assumption more realistic and dampening the selection bias. Of course, it is not always possible to do so when the non-probability sample is not managed by the same agency as the probability sample. In some cases, there may be many auxiliary variables available for adjusting estimates from the non-probability sample. Variable selection techniques can be useful for selecting relevant auxiliary variables but need to be adapted to this data integration context. This problem is currently being investigated at Statistics Canada.

Quality indicators for estimates based on integrated data is a topic that requires further research. The model variance of those estimators can be estimated, often using standard techniques, but the resulting variance estimates fail to account for the selection bias, which may not be negligible. Sensitivity analysis may provide a useful complement to variance estimates. It allows for assessing the effect of omitting a relevant auxiliary variable on the conclusions drawn from the available data.

Ding and VanderWeele (2016) is a recent paper on the topic that is currently being investigated at Statistics Canada.

In some cases, depending on the users' objectives, evidence will suggest that estimates from integrated data are not appropriate for the intended uses. Still, it might be desired to take advantage as much as possible of the available non-probability sample data. Small Area Estimation is one possible avenue that combines probability and non-probability sample data through models. Although the Fay-Herriot model discussed in Section 6 relies on the validity of model assumptions, the resulting inferences are not as dependent on the model as inferences based on methods discussed in Section 5. First, diagnostics can be used to assess the model adequacy (e.g., Hidiroglou, Beaumont and Yung, 2019). Also, when the probability sample size is large in a domain, the small area estimates are usually quite close to the reliable direct estimates. However, SAE is only feasible when the variable of interest is observed in the probability sample. Another alternative to make use of non-probability sample data is to use dual frame weighting methods (e.g., Kim and Tam, 2020; Rao, 2020; Beaumont, 2020). The advantage of this approach is that it remains design-based, thereby not dependent on any model assumptions.

We have not discussed other practical issues related to the use of big data and non-probability samples, such as privacy, access, and transparency. A report by the U. S. National Academies of Sciences, Engineering and Medicine (2017) extensively treats the privacy issue, in addition to methodology for integrating data from multiple sources.

Acknowledgements

We deeply thank Andrew Brennan and Joanne Charlebois from Statistics Canada for providing the two figures in section 4, and David Haziza for his careful reading of the paper.

References

Baribeau, B. (2020). Trial by COVID for Statistics Canada's web panel pilot. Internal document, Statistics Canada.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, **46**, 1-28.

Bethlehem, J. (2020). Working with response probabilities. *Journal of Official Statistics,* **36**, 647-674.

Biemer, P. (2019). Can a survey sample of 6000 records produce more accurate estimates than an administrative data base of 100 million? (The answer may surprise you). *The Survey Statistician*, **80**, 11-15.

Chatrchi, G., Gauvin, H., Ding, A. (2020). Combining survey data with other data sources using small area estimation: a case study. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, **104**, 439-453.

Chen, Y., Li, P., and Wu, C. (2019). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* (published online).

Ding, P., and VanderWeele, T.J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, 27, 368-377.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **85**, 398-409.

Hidiroglou, M.A., Beaumont, J.-F. and Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, **45**, 1, 101-126. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019001/article/00009-eng.pdf.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**, 375-394.

Kim, J.K., and Tam, S.M. (2020). Data integration by combining big data and survey data for finite population inference. Unpublished manuscript.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox and the 2016 US presidential election. *Annals of Applied Statistics*, **12**, 685-726.

National Academies of Sciences, Engineering, and Medicine (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection*: *Nest Steps.* The National Academies Press, Washington, DC. https://doi.org/10.17226/24893.

Neyman, J. (1934). On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society*, **97**, 558--625.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, published online April 2020.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Second Edition, Hoboken, New Jersey: John Wiley & Sons, Inc.

Renaud, M., and Beaumont, J.-F. (2020). Crowdsourcing during a pandemic: the Statistics Canada experience. Paper presented at the Advisory Committee on Statistical Methods, Statistics Canada, October 27, 2020.

Rivers, D. (2007). Sampling for web surveys. In: *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Van den Brakel, J.A. and J. Bethlehem (2008), Model-based estimation in official statistics. CBS Discussion paper 08002. Statistics Netherlands, The Hague/Heerlen. https://www.cbs.nl/nl-nl/achtergrond/2008/10/model-based-estimation-for-official-statistic.

Yang, S. and Kim, J. (2020). Statistical data integration in survey sampling: a review. *arXiv preprint arXiv:2001.03259v1.*

The Survey Statistician, 2021, Vol. 83, 23-26.

Official Statistics at the Crossroads: Data Quality and Access in an Era of Heightened Privacy Risk[1]

John M. Abowd[2]

[2] Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau, john.maron.abowd@census.gov

Abstract

This paper discusses the challenging problem of balancing the competing interests of access to high-quality statistical data and privacy protection. The paper argues that an optimal choice must account for the preferences of data users and providers, and that technology developed by cryptographers, such as differential privacy, may help us find efficient algorithms for implementing such an optimal choice.

Keywords: confidentiality, privacy, data quality, data access

There are many stakeholders in the conversation about the appropriate response to heightened privacy risk. In the United States, the American Statistical Association, the Committee on National Statistics (a standing committee of the National Academy of Sciences), and the Chief Statistician of the U.S. (in the Office of Management and Budget), have all furthered the discussions of privacy protection. Statistical agency experts and directors that I have briefed over the course of the last four years, since I took this job at the Census Bureau, have also provided their frank and illuminating discussion of the issues that they face in attempting to modernize their disclosure avoidance methods.

Data users, including those from academia, research organizations, industry, and inside the Census Bureau, have also joined the discussion. Serious data users understand why it is worth multiple billions of dollars to conduct a population census. They know that conducting a quality census is more than just a statement by the statistical agency for public relations. They know that census data in the U.S. allocate at least six hundred and seventy-five billion dollars of federal funds every year with some estimates placing the annual value closer to one and a half trillion dollars. They know that the U.S. House of Representatives is apportioned based on data from the census, reminding us that it is important to conduct our censuses and surveys in a manner that maintains the integrity of those data and that allows the user community to have faith in their fitness for use. This is part of the U.S. Census Bureau's dual mandate.

The other part of that dual mandate is to protect the confidentiality of the respondents and the data that they provided. This is a challenging problem, and we all have to acknowledge that both sides of this discussion – really the entire continuum within this discussion – bring legitimate viewpoints to the table that must be respected and considered before making final decisions. It is important for me to acknowledge that, and I think most statisticians and researchers reading this article would not find anything controversial in that acknowledgement.

Traditional statistical disclosure limitation is broken. That does not mean it usually fails. It does not usually fail. It does have significant vulnerabilities exposed by the cryptographers who migrated from

[1] Based on a talk originally delivered at the Joint Statistical Meetings, Denver, CO, USA on July 30, 2019. The views expressed in this article are those of the author and not the U.S. Census Bureau.

computer science into safe data publication. We must now address those vulnerabilities. They are real. They are documented in thousands of carefully-prepared, well peer-reviewed scientific papers that explain what the vulnerabilities mean and why the traditional methods are so exposed—not in an extreme sense, but vulnerable in precisely the sense that computer scientists have properly defined.

There is a very steep learning curve for the official statistics community to climb, because these methods come from a different scientific tradition and involve very different methodologies. Extremely talented mathematical statisticians – well versed in the theory that underlies our statistical analyses – in particular, estimation based on complex multi-stage probability samples – were not exposed to the mathematical reasoning that underlies differential privacy and formal privacy systems. That is just a fact – not a statement of incompetence on anyone's part. My own mathematical background also did not include most of the tools necessary to understand the arguments that the cryptographers were making. But, we do need to face up to those vulnerabilities – we need to rethink how we approach confidentiality protection, and we need to do it so that our future disclosure limitation systems can deliver the same promise of quality and confidentiality protection that they delivered when they were originally conceptualized, primarily by mathematical statisticians in the 1970s. So now, let's dive into it.

Privacy protection is an economic problem. It is not a computer science problem; it is not a statistics problem; it is an *economic problem*. It is about the allocation of a scarce resource – the information in the confidential data that statistical agencies collected – between two competing uses: public data products and privacy protection. The confidential data are a scarce resource because they are finite. If finite, that must mean that the published data products can fully consume the information leaving no privacy protection, if we are not careful. That is precisely what the economic analysis of confidentiality protection teaches us: computer science informs the technology for transforming confidential data into useful information products, and making the publication algorithms produce accurate, fit-for-use information products that also protect confidentiality is a function of using good computer science.

Computer science has thus defined the production possibility frontier between privacy protection and accuracy of publications just the same way as the toy examples in your Introductory Microeconomics class defined the guns and butter production possibility frontier: if you consume more guns, you will have less butter. If we consume more accuracy, we will have less privacy – that is a mathematical fact. If we do it carelessly, then we will inefficiently spend privacy-loss, as I prefer to call it, and not get as much accuracy as we could. If we do it carefully, then we will use algorithms that are on that production possibility frontier – algorithms that are efficient. But if you claim that you can get more accuracy than an efficient privacy-enhancing data publishing technique, you are claiming something that is mathematically false. The comparable claim that traditional statistical disclosure limitation can be more accurate and just as privacy-preserving is also mathematically false. The traditional methods are Pareto dominated, in the economic sense, by the formally private methods, of which differential privacy is the leading example. That means the traditional methods are on the interior of the production possibility frontier – you can improve the accuracy, reduce the privacy loss or both at no cost by moving to the efficient frontier. The technology that cryptographers brought to data publication describes the efficient frontier. It is not constant. It changes every day with new research. This research can, and does, make the algorithms more efficient. It pushes the production possibility frontier outwards. That is the technology side.

What does it mean to have an optimal balance of accuracy relative to privacy protection? Answering this question requires more than technology. An optimal choice must account for the preferences of data users and providers. It must summarize the extraordinarily heterogeneous costs to the providers in terms of their privacy loss and the equally heterogeneous benefits to the users in terms of the accuracy of the publications for their intended uses. In short, we have to balance the social preferences of data users and providers to determine an optimal trade-off.

I want you to look yourself in the mirror now, because I have. How much weight do you personally put on data accuracy versus privacy protection? In your way of thinking about the world, whose job is it to protect the privacy interests of the data contributors? I think most official statisticians would say that is our job. Whose job is it to protect the fitness for use of the data products that we release? Again, I think most official statisticians would say that is also our job. Those are *competing* interests, and so it must be our job to balance those competing interests. We do not have a complete repertoire of tools with which to perform that job. We need to fix that, and that is a critical part of our research mission.

I want you to ask yourself this question. If Facebook said "*If you think you have re-identified someone in public data that we released for research purposes, you can't be sure that you are correct because we used disclosure limitation techniques for which we cannot give you the details,*" what would you say?

I also want you to ask yourself this question. If you were refereeing a scientific paper, and the author said "*My inferences may not be valid, because the agency that provided data access did not release details sufficient to correct for bias and variability due to statistical disclosure limitation,*" what would you say to the editor?

I think I can guess the answer to both of those questions. We have to find a way out of this situation. Statistical agencies should not behave in a manner that we would find unacceptable for Facebook or scientific journals. Official statisticians cannot continue to ask the users of our data to ignore the things that we have to do to protect confidentiality. We need to provide data analysis systems that are statistically valid – systems where the inferences are correct according the appropriate underlying mathematical theory, just as we did for design-based estimation from probability samples. We must be able to say that we protected the confidentiality of these data using these algorithms with these parameters. Statistical disclosure limitation should not remain a back-room operation exempt from the scrutiny of either the providers or the users. Users and providers must assess the quality of our confidentiality protections. National statistical agencies should correct, strengthen, or otherwise adjust their publication systems based on the reasoned analyses of both data and privacy advocates. Speaking only for myself, our research and methodology ought to step up to this challenge.

Selected references

The database reconstruction vulnerability:

Dinur, I. and Nissi, K. (2003) Revealing information while preserving privacy. In: *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03)*. ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.

Differential privacy:

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis." In: *Theory of Cryptography Conference* (eds. S. Halevi and T. Rabin), pp. 265-284. Springer Berlin Heidelberg. DOI: 10.1007/11681878_14.

Dwork, C. (2006) Differential Privacy, *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, Springer Verlag, 4052, 1-12, ISBN: 3-540-35907-9.

Dwork, C. and Roth, A. (2014) The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, **9**, 211–407, DOI: 10.1561/0400000042.

Economic analysis:

Abowd, J.M. and Schmutte, I.M. (2019) An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, **109**, 171-202, DOI:10.1257/aer.20170627. [AER, ArXiv preprint, Replication information]

Abowd, J.M., Schmutte, I.M., Sexton, W. and Vilhuber. L. (2019) Why the economics profession must actively participate in the privacy protection debate. *American Economic Association: Papers and Proceedings*, **109,** 397-402, DOI:10.1257/pandp.20191106. [download preprint].

Kifer, D. and Machanavajjhala, A. (2011) No free lunch in data privacy. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11)*. ACM, New York, NY, USA, 193-204. DOI:10.1145/1989323.1989345.

More information on the disclosure limitation system for the 2020 Census of Population and Housing in the United States:

Main U.S. Census Bureau web page on statistical disclosure limitation for the 2020 Census https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

International Conference on Machine Learning keynote address: Abowd, J.M. (2019) The Census Bureau tries to be a good data Steward in the 21st century. *International Conference on Machine Learning (ICML)*, keynote address. [video, start at minute 18:00]

The Survey Statistician, 2021, Vol. 83, 27–37.

New and Emerging Methods

Graph sampling: An introduction

Li-Chun Zhang[1,2,3]

[1]University of Southampton, UK, L.Zhang@soton.ac.uk
[2]Statistics Norway, Norway
[3]University of Oslo, Norway

Abstract

Representing a collection of relevant units by a graph allows one to incorporate the connections (or links) among the units in addition to the units themselves. One may be interested in the structure of the connections, or the links may provide effectively access to those units that are the primary interest. Either way, graph sampling provides a statistical approach to study real graphs. Just like sampling from finite populations, it is based on exploring the variation over all possible *sample graphs*, which can be taken from the given *population graph* according to a specified method of probability sampling, and design-based inference using a suitable graph sampling strategy is valid “whatever the unknown properties” (Neyman, 1934) of the population graph.

Keywords: graph, probability sampling, observation procedure, motif of interest, sampling strategy

1 Sampling from real graphs

Birnbaum and Sirken (1965) consider ‘indirect sampling’ of patients via an initial sample of medical centres. Since any in-scope patient may receive treatment at multiple places, not all of which are among the actual sample of medical centres, additional knowledge of all the out-of-sample treatment places of each sampled patient needs to be collected in order to calculate the patient’s sample inclusion probability. In addition to the Horvitz-Thompson estimator (HTE, Horvitz and Thompson, 1952), Birnbaum and Sirken (1965) propose 2 unbiased estimators in this unusual situation.

Zhang (2020b) considers sampling to estimate the prevalence of an epidemic in a given population U, where one would like to increase the sample yield of *cases*, i.e. persons with $y_i = 1$ in contrast to *noncases* with $y_i = 0$, in order to improve the design efficiency. Let s_0 be an initial sample from U, with inclusion probability $\pi_i = \Pr(i \in s_0)$. Since the virus is transmitted via personal contacts, consider *adaptive network tracing*, where all the contacts of each case i in s_0 are included, and the procedure is repeated for them, and so on until no more cases can be added in this way. Let $\pi_{(i)} = \Pr(i \in s)$, where s is the final sample. Since any case i that is in contact with other cases can be included in s

by adaptive network tracing, even when it is not selected in s_0 initially, we achieve $\pi_{(i)} \geq \pi_i$. This is a special case of adaptive cluster sampling (ACS, Thompson, 1990) with binary y_i.

Zhang and Oguz-Alper (2020) develop the theory, which enables one to represent both the situations above as sampling from a *bipartite incidence graph (BIG)*. Patone and Zhang (2020) develop generally the *incidence weighting estimator (IWE)* under BIG sampling (BIGS), which encompasses all the estimators considered by Birnbaum and Sirken (1965). BIGS and the associated IWE form a flexible *graph sampling strategy*, which extends the finite-population (FP) sampling strategy consisting of a probability sampling design and the associated HTE. The BIGS-IWE strategy is applicable to many unconventional probability sampling techniques, which "are not explicitly stated as graph problems but which can be given such formulations" (Frank, 1977), including indirect sampling (Birnbaum and Sirken 1965; Lavalleè, 2007), network sampling (Sirken, 1970; 2005), adaptive cluster sampling (Thompson, 1990, 1991) and line-intercept sampling (Becker, 1992; Thompson, 2012). See Zhang and Oguz-Alper (2020) and Patone and Zhang (2020) for the relevant discussions.

As Zhang and Patone (2017) point out, in all the aforementioned situations, one is interested — rather conventionally — in some finite population total (or mean), where the connections (or links) among the relevant population elements and sampling units — more or less unconventionally — provide the access to the target population, which otherwise would have been ineffective or impractical to sample. Meanwhile, in sampling from arbitrary graphs generally, one is typically interested in the structure of the links themselves, often expressed in terms of a particular *motif*, which may simply be defined as a subgraph of specific characteristics. An early example is snowball sampling by Goodman (1961), where the motif of interest is 'pair with mutual relationships' in a special graph where all the nodes have out-degree one. In a series of work spanning over several decades (e.g. Frank, 1971, 1977, 1978, 1979, 1980, 1981, 2011), Ove Frank studies from this perspective graph sampling of motifs defined for nodes, dyads, triads (star, triangle), components, etc. Zhang and Patone (2017) provide a structure of *graph totals* of various motifs, to reflect the extended scope of investigation.

Thus taken together, representing a population of relevant units by a graph allows one to incorporate the connections (or links) among the units in addition to the units themselves. One may be either interested in the characteristics of the graph, or the links may provide effectively access to those units that are the primary interest. Either way, graph sampling provides a statistical approach to study real graphs. Just like sampling from finite populations, it is based on exploring the variation over all possible *sample graphs*, which can be taken from the given *population graph* according to a specified method of probability sampling, and design-based inference using a suitable graph sampling strategy is valid "whatever the unknown properties" (Neyman, 1934) of the population graph.

As much as graph sampling is versatile, it can be intricate when it comes to the formulation of graph sampling strategy in various situations. Below are three key elements in any case.

I. Definition of sample graph. Zhang and Patone (2017) define sample graph, where the specified sample observation procedure makes use of incident edges. Other observation procedures are conceivable which, in particular, may involve random jumps or teleporting to non-adjacent nodes. Tweaks of the definition of sample graph are needed accordingly.

II. Basis of inference. Zhang and Patone (2017) synthesise the existing graph sampling theory, where inference is based on the sample graph inclusion probabilities of the motif of interest. The IWE makes more extensive use of the same basis of inference, allowing for many unbiased estimators in addition to the HTE. More generally, inference can be based on other avaiable *sampling probabilities* associated with the given graph sampling method, as e.g. will be discussed for random walk sampling, which call for principally different strategies.

III. Eligible sample motifs. A motif that is observed in the sample graph is nevertheless 'ineligible' for estimation, if the required probabilities for inference cannot be calculated. Eligibility of a particular sample motif depends on the availability of the knowledge of its *ancestry* (Zhang and Patone, 2017). Essentially, apart form the actual way by which a motif is sampled, one needs to know how else it could have been sampled under the given sampling method. The concept of ancestry under graph sampling generalises the concept of *multiplicity* defined by Birnbaum and Sirken (1965), where it amounts to the knowledge of the out-of-sample medical centres for each sampled patient. Identification of eligible sample motifs is the key to any feasible graph sampling strategy.

In the rest of the paper, examples will be given to elaborate the points above. For the details that may be necessary for a fuller comprehension the reader is kindly referred to the relevant sources.

2 BIGS-IWE generalises FP-sampling and HT-estimation

Denote by $\mathcal{B} = (F, \Omega; H)$ a population BIG, where the node set is bipartitioned into F and Ω, such that (directed) edges exist only from F to Ω, denoted by $(i\kappa) \in H$, iff the selection of $i \in s_0 \subset F$ leads to that of κ from Ω. As explained and illustrated below, the strategy BIGS-IWE generalises the familiar strategy of 'FP-sampling and HT-estimation'.

Denote by $U = \{1, ..., N\}$ a *population* of size N. Let y_k be a constant associated with each $k \in U$, with population total $\theta = \sum_{k \in U} y_k$. Denote by s a sample from U, according to a method of probability sampling, where the sample inclusion probability $\Pr(k \in s)$ is either known in advance or can be calculated for the sample units afterwards. The HTE of θ is $\hat{\theta}_{HT} = \sum_{k \in s} y_k / \Pr(k \in s)$.

For element sampling, let $F = \Omega = U$, where $(i\kappa) \in H$ iff i and κ refer to the same population element. The correspond BIGS representation is given to the left in Figure 1. For cluster sampling, illustrated to the right in Figure 1, let F consist of the clusters (of which there are M in total) and $\Omega = U$ the elements that are nested in the clusters, where $(i\kappa) \in H$ iff element κ belongs to cluster i. Clearly, the strategy BIGS-HTE suffices for these familiar FP-sampling situations.

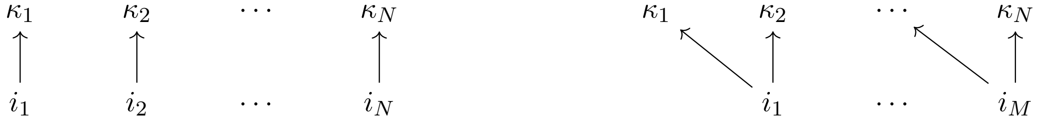


Figure 1: BIGS representation of finite-population sampling of elements (left) or clusters (right).

For indirect sampling of Birnbaum and Sirken (1965), let F consist of the medical centres and Ω the patients of interest, where $(i\kappa) \in H$ iff patient κ receives treatment at centre i. The BIG is illustrated in Figure 2, where the mapping from F to Ω can be many-many, instead of simply one-one or one-many as in Figure 1. Let us consider the elements I - III, in order to arrive at the strategy BIGS-IWE.

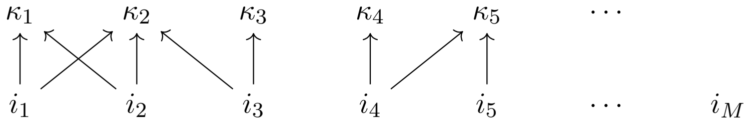


Figure 2: BIGS in general

I. Following the definition of Zhang and Patone (2017), the sample graph under BIGS from $\mathcal{B}$ is generally given as follows. Let s_0 be an initial sample from F, with sample inclusion probabilities π_i, π_{ij}, etc. Apply the *incident forward observation procedure*, by which all the out-edges from s_0 are included in the sample of edges, denoted by $H_s = \{(i\kappa) : (i\kappa) \in H, i \in s_0\}$. The sample nodes are the union of s_0 and those incident to the edges in H_s, i.e. $s_0 \cup \Omega_s$, where $\Omega_s = \alpha(s_0) = \cup_{i \in s_0} \alpha_i$, and

$\alpha_i = \{\kappa : (i\kappa) \in H\}$ are the *successors* of i in $\mathcal{B}$. The sample graph under BIGS is given by

$$\mathcal{B}_s = (s_0, \Omega_s; H_s).$$

II. The inference is based on the sample inclusion probabilities. The fact that each κ in Ω_s can possibly be accessed via multiple sampling units in F calls for the concept of ancestry under graph sampling. For instance, suppose $i_1 \in s_0$ but not i_2 or i_3. To calculate the inclusion probability of the patient $\kappa_2 \in \alpha_{i_1}$, one must collect the information that it receives treatment at i_2 and i_3 as well, i.e. how else it could have been sampled under BIGS here.

III. The ancestry knowledge of $\kappa \in \Omega_s$ is secured and it is eligible for estimation of $\theta = \sum_{\kappa\in\Omega} y_\kappa$, where y_κ is a constant associated with each $\kappa \in \Omega$, provided the observation of $\beta_\kappa \setminus s_0$, where $\beta_\kappa = \{i \in F : (i\kappa) \in H\}$ are the *predecessors* of κ in $\mathcal{B}$, although $\beta_\kappa \setminus s_0$ are not part of the sample graph $\mathcal{B}_s$. It follows that all the nodes in Ω_s are eligible, provided the observation of $\beta\big(\alpha(s_0)\big) \setminus s_0$ in addition to $\mathcal{B}_s$, where $\beta\big(\alpha(s_0)\big) = \cup_{\kappa\in\alpha(s_0)}\beta_\kappa$.

Let $W_{i\kappa}$ be the *incidence weight* associated with each edge $(i\kappa) \in H_s$. The IWE of θ is given by

$$\hat{\theta} = \sum_{(i\kappa)\in H_s} W_{i\kappa} \frac{y_\kappa}{\pi_i} \tag{1}$$

(Patone and Zhang, 2020), where $\pi_i = \Pr(i \in s_0)$ is also the probability that $(i\kappa)$ is included in $\mathcal{B}_s$ under BIGS. While the HTE is defined for Ω_s, the IWE is defined for the sample BIG edge set $H(\mathcal{B}_s) = H_s$, where each $(i\kappa)$ in H_s is incident to i in s_0 and κ in Ω_s. Patone and Zhang (2020) show that the IWE encompasses all the estimators considered by Birnbaum and Sirken (1965). In particular, the HTE is a special case, where $W_{i\kappa}$ varies according to $s_\kappa = s_0 \cap \beta_\kappa$, subjected to the condition that ensures unbiased estimation of θ over repeated sampling: for any $\kappa \in \Omega$,

$$\sum_{i\in\beta_\kappa} E(W_{i\kappa} | i \in s_0) = 1.$$

This generalises the result of Birnbaum and Sirken (1965) for constant weights, denoted by $\omega_{i\kappa}$ for distinction, which requires $\sum_{i\in\beta_\kappa} \omega_{i\kappa} = 1$ for any $\kappa \in \Omega$, including $\omega_{i\kappa} = 1/m_\kappa$ and $m_\kappa = |\beta_\kappa|$.

3 BIGS-IWE for unconventional sampling: ACS as an example

3.1 ACS with binary outcome variable

Let U be the population of size N and $\mu = \theta/N$ the prevalence of interest. Adaptive network tracing requires the population U to be represented by the population graph $G = (U, A)$ where, in addition to the node set U, the edge set A contains all the relevant contacts. We shall treat the graph as undirected and simple, where $(ij), (ji) \in A$ if persons i and j are in-contact, and there is only one edge either way regardless of the frequency or intensity of the contact between i and j.

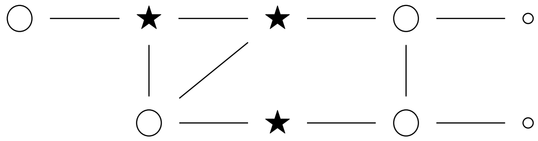


Figure 3: Cases ★, noncase edge nodes ◯, other noncases ∘

Figure 3 illustrates a part of such a population graph G, with stars for cases and circles for noncases. In particular, all the cases are partitioned into *case networks*, where those in the same network all have $y_i = 1$ and are connected to each other in G; two case networks are shown in Figure 3. Next, a (network) *edge node* is a noncase that is adjacent to at least one case network; the four edges nodes in Figure 3 are shown as bigger circles than the other noncases.

ACS from G employs contact tracing starting from an initial sample s_0 from U, which is adaptive because tracing is only applied to the contacts of ★ but not ◯ or ∘. The final sample s by ACS can be divided into three parts: (i) a set of case networks, (ii) the edge nodes, and (iii) the remaining noncase nodes in the initial sample s_0 which do not belong to (i) and (ii). Zhang (2020b) considers the efficacy of several ACS designs for cross-sectional as well as change estimation of prevalence. These graph sampling methods allow one to unite tracing for combating the disease *and* sampling for estimating the prevalence during an epidemic outbreak.

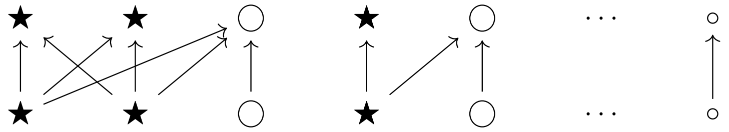


Figure 4: BIGS representation of ACS from G with binary outcome variable

Following Zhang and Oguz-Alper (2020), one can apply the BIGS-IWE strategy to ACS from G above, thereby allow other unbiased estimators in addition to the HTE. Let $F = \Omega = U$ in $\mathcal{B}$, and let H contain the incident observation relationships among the nodes in G under adaptive network tracing. For instance, let the two leftmost ★ in Figure 4 be the two in the same network in Figure 3, and let ◯ next to them in Figure 4 be one of their edge nodes in Figure 3. Under ACS from G, the selection of either ★ in s_0 leads to all the three of them to be included in the sample s, yielding the corresponding edges from these two ★ in $\mathcal{B}$. Meanwhile, selecting any ◯ in s_0 does not lead on to any adjacent case network, such that ◯ has only an edge to itself in $\mathcal{B}$. Similarly, the other two edge nodes in Figure 3 can be included in $\mathcal{B}$, which are omitted here to avoid cluttering the figure visually.

Thus, the sample graph under BIGS from $\mathcal{B}$ is the same as that by ACS from G. The inference basis is still the relevant sample inclusion probabilities. Since each sample case is observed together with its network under ACS, the knowledge of its ancestry is secured for the BIGS representation with $\mathcal{B}$ defined above. But ancestry is generally unclear for any edge node ◯ in Figure 4, since we would not observe any of its ★-ancestors in a case network unless that network happens to intersect s_0. *However, this does not matter here, since a noncase κ in Ω with $y_\kappa \equiv 0$ contributes nothing to the IWE* (1) *regardless its inclusion probability.* Thus, using only the sample case nodes as the eligible motifs in Ω_s, the BIGS-IWE is a feasible strategy for ACS from G with binary outcome variable.

3.2 ACS with continuous outcome variable

Simply ignoring the edge nodes would not be valid for ACS with continuous outcome variable, where an edge node generally has a non-zero value below the threshold chosen for adaptive sampling. Thompson (1990) proposes an inferential approach, where one modifies two of the estimators of Birnbaum and Sirken (1965). Zhang and Oguz-Alper (2020) develop the BIGS-IWE strategy. Let us illustrate their approach here using the example of Thompson (1990).

The population U consists of $N = 5$ spatial grids, with associated $y_U = \{1, 0, 2, 10, 1000\}$ for the amount of species of interest. Each grid has either one or two neighbours which are adjacent in the undirected graph $G = (U, A)$ below, where we simply denote each grid (or node) by its y-value as Thompson (1990). This is a *valued* graph where G is known but the associated y_U are unknown.

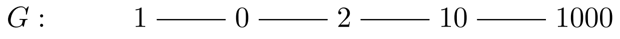


Figure 5: Graph for ACS (Thompson, 1990)

Given an initial sample s_0 of size 2 by simple random sampling (SRS) from U, one would survey all the adjacent grids (in both directions if possible) of a sample grid i if y_i exceeds the threshold value 5 but not otherwise, and so on. A network in G may consist of one or more connected nodes all with y-values above the threshold such as $\{10, 1000\}$ here, or it may be a single node with y-value up to the threshold, some of which are edge nodes such as node 2 here. The interest is to estimate the mean amount of species per node, denoted by $\mu = \theta/N$, where $\theta = \sum_{i \in U} y_i$.

Since the sample inclusion probability of any edge node is generally unknown under ACS from G, Thompson proposes to modify the HTE, such that an edge node i is used for estimation (i.e. eligible) only when $i \in s_0$ directly, the probability of which is $\pi_i = \Pr(i \in s_0) = n/N$ under SRS of s_0. Similar modification can be applied to the 2nd estimator of Brinbaum and Sirken (1965), which is referred to as the *Hansen-Hurvitz (HH) type* estimator by Thompson (1990).

Zhang and Oguz-Alper (2020) denote the strategy of Thompson (1990) by $(\mathcal{B}, \hat{\theta}^*_{HT})$ when the modified HTE is used as the estimator, where the population $\mathcal{B}$ has $F = \Omega = U$ and the edge set H contains all the observational links under ACS from G. They observe that it is as well possible to modify the sampling when constructing a feasible strategy, say, (ACS*, HT) or (ACS*, HH). In particular, they use BIGS as ACS*, in which case the IWE would unify and generalise the HTE and HH-type estimator.

For a generally feasible strategy with BIGS one can use instead $\mathcal{B}^* = (U, U; H^*)$ in Figure 6. The observational links $(10, 2)$ and $(1000, 2)$ under ACS from G are removed to ensure ancestral observation in $\mathcal{B}^*$. For instance, given $s_0 = \{0, 2\}$, the observation procedure of ACS means 10 and 1000 are not observed, as in $\mathcal{B}^*$ where 2 in $\Omega(\mathcal{B}^*)$ has only itself as the ancestor in $F(\mathcal{B}^*)$. One can now use the unmodified HTE under BIGS from $\mathcal{B}^*$, as a special case of IWE, denoted by $(\mathcal{B}^*, \hat{\theta}_y)$.

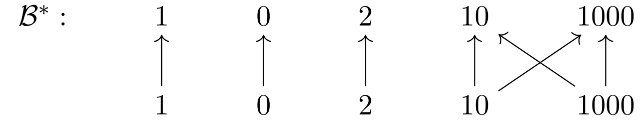


Figure 6: A feasible BIGS-IWE strategy for ACS

The two strategies $(\mathcal{B}, \hat{\theta}^*_{HT})$ and $(\mathcal{B}^*, \hat{\theta}_y)$ lead to the same estimator, since the eligible sample nodes in Ω_s are the same under both. The difference is that applying the Rao-Blackwell method does not change $\hat{\theta}_y$ under BIGS from $\mathcal{B}^*$, whereas it changes $\hat{\theta}^*_{HT}$ under BIGS from $\mathcal{B}$.

Another possible strategy using BIGS in this particular setting is to make an edge node ineligible, if itself is selected in s_0 but not its neighbouring above-threshold network, with $\mathcal{B}^\dagger$ in Figure 7. Denote this strategy by $(\mathcal{B}^\dagger, \hat{\theta}_y)$. It is feasible here because the egde node 2 has only one above-threshold neighbouring network in G, i.e. $\{10, 1000\}$; but it would be infeasible generally provided an edge node has two or more such networks in G.

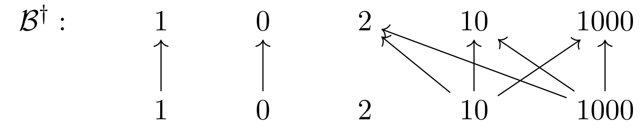


Figure 7: Another BIGS-IWE strategy for ACS

The BIGS-IWE strategy extends the inferential approach to ACS pioneered by Thompson (1990), where one can modify *either* part of a standard strategy (sampling, estimator) when it is otherwise infeasible in a given situation. For the example discussed above, Table 1 taken from Zhang and Oguz-Alper (2020) provides the numerical details of the three strategies $(\mathcal{B}, \hat{\theta}^*_{HT})$, $(\mathcal{B}^*, \hat{\theta}_y)$ and $(\mathcal{B}^\dagger, \hat{\theta}_y)$.

Table 1: Strategies using BIGS for ACS from $G:\ 1-0-2-10-1000$.

	$(\mathcal{B},\hat{\theta}^*_{HT})$		$(\mathcal{B}^*,\hat{\theta}_y)$		$(\mathcal{B}^\dagger,\hat{\theta}_y)$	
s_0	Ω_s	$\hat{\mu}^*_{HT}$	Ω_s	$\hat{\mu}_y$	Ω_s	$\hat{\mu}_y$
1,0	1,0	0.500	1,0	0.500	1,0	0.500
1,2	1,2	1.500	1,2	1.500	1	0.500
0,2	0,2	1.000	0,2	1.000	0	0.000
1,10	1,10,*2*,1000	289.071	1,10,1000	289.071	1,10,2,1000	289.643
1,1000	1,1000,*2*,10	289.071	1,1000,10	289.071	1,1000,2,10	289.643
0,10	0,10,*2*,1000	288.571	0,10,1000	288.571	0,10,2,1000	289.143
0,1000	0,1000,*2*,10	288.571	0,1000,10	288.571	0,1000,2,10	289.143
2,10	2,10,1000	289.571	2,10,1000	289.571	2,10,1000	289.143
2,1000	2,1000,10	289.571	2,1000,10	289.571	2,1000,10	289.143
10,1000	10,1000,*2*	288.571	10,1000	288.571	10,1000,2	289.143
Variance		17418.4		17418.4		17533.7

4 Graph sampling: On to general theory

Sampling from arbitrary graph $G = (U, A)$ involves a number of conceptual generalisations of BIG sampling discussed above. Given limited space, we focus our discussion here on the following.

- So far we have only seen examples of motifs defined for the nodes of G. Zhang and Patone (2017) define generally motifs and their graph total, i.e. instead of population total over U.
- There are many observation procedures in a graph, which make use of the edges incident to an initial sample of nodes. Zhang and Patone (2017) discuss induced and incident procedures.
- Repeating an incident observation procedure leads to *multiwave* sampling. *T-wave snowball sampling* (T-SBS, Zhang and Patone, 2017) is the probabilistic version of breath-first search in graphs, and *targeted random walk (TRW)* is that of depth-first search.
- Zhang and Patone (2017) consider inference based on the graph sample inclusion probabilities. Depending on the observation procedure, other sampling probabilities may be necessary.

Definitions Let $G = (U, A)$ consist of nodes U and edges A. Let A_{ij} contain the edges from i to j, and $a_{ij} = |A_{ij}|$. Attaching values to U or A yields a *valued graph*. One may consider a graph to be the *structure* of a valued graph. We do not consider separately sampling from graphs or value graphs. Generally speaking, a graph sampling method may depend on the values associated with G, and the values associated with the sample graph G_s are observed together with G_s.

Let M be a subset of U. Let $G(M)$ be the subgraph *induced* by M, whose edge set is given by $\{A_{ij} : (i,j) \in M\}$. A subgraph $G(M)$ with specific characteristics is called a *motif*, denoted by $[M]$. For example, $[i : a_{i+} = 3]$ is a motif of node with out-degree 3, $[i, j : a_{ij}a_{ji} = 1]$ of a node pair with mutual simple relationship, and $[i, j : a_{ij} + a_{ji} = 0]$ of a non-adjacent node pair.

Let $y\big(G(M)\big)$, or simply $y(M)$, be a function of $G(M)$. Let Ω contain all the relevant M. Let

$$\theta = \sum_{M\in\Omega} y(M) \tag{2}$$

be the *graph total* over Ω. It is said to be the *k-th order*, if $|M| = k$ for all $M \in \Omega$. Although It is possible to let Ω in (2) be the set of motifs of interest directly, and let y_κ be a function of motif κ, it can be convenient if the summation is over all the relevant node sets. For instance, the motifs $[i, j : a_{ij}a_{ji} = 1]$ can be enumerated over $\Omega = \{(i, j) : i \neq j \in U\}$, with $y(i, j)$ as the corresponding counter. If Ω is the set of these motifs, then writing $\theta = |\Omega|$ is more natural than $\theta = \sum_{\kappa\in\Omega} 1$.

Zhang and Patone (2017) consider induced or incident *observation procedure* given an initial sample of nodes, denoted by s_0. Take $G:\ a \quad b \to c \to d$ for example. Let $s_0 = \{b, d\}$. We observe none of the edges if the observation procedure is induced, or the edge (bc) if it is incident forward, or (cd) if incident backward, or both (bc) and (cd) if incident reciprocal.

Ove Frank studies sampling of node sets or motifs using such observation procedures, where a sample of motifs from the population of motifs is conceived in analogy to a sample s from the population U. Zhang and Patone (2017) define the *sample graph* $G_s = (U_s, A_s)$ as a subgraph of G.

- Initial sample of nodes $s_0 \subset U$, with $p(s_0)$, π_i, π_{ij}, etc.
- Application of the specified observation procedure, starting from s_0.
- Specify the reference set $s_{\text{ref}} \subset U \times U$, such that $A_s = A \cap s_{\text{ref}}$
- Let $U_s = s_0 \cup \text{Inc}(A_s)$, where $\text{Inc}(A_s)$ denotes the nodes incident to the edges in A_s.

Compared to sampling from finite populations, a defining feature of sampling from graphs is that one uses the edges. The definition of sample graph above includes the situation, where the initial node sample s_0 is given as the nodes that are incident to a sample of edges directly selected from A. Direct sampling of edges may be useful e.g. if G is known but is too large to be counted. It is then possible for the observation procedure to specify that no additional edges need to be sampled.

It is convenient to specify the sample edges A_s via s_{ref}, which explicates the parts of the adjacency matrix $[a_{ij}]$ that are observed given s_0 and the observation procedure. Take again $G:\ a \quad b \to c \to d$ for example. Let the rows and columns of $[a_{ij}]$ be arranged in the order a, b, c, d. The set s_{ref} given $s_0 = \{b, d\}$ and the various observation procedures are shown in $\boxed{\text{1/0}}$ below.

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \boxed{0} \\ 0 & 0 & 0 & 1 \\ 0 & \boxed{0} & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 0 \\ \boxed{0} & \boxed{0} & \boxed{1} & \boxed{0} \\ 0 & 0 & 0 & 1 \\ \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} \end{bmatrix} \quad \begin{bmatrix} 0 & \boxed{0} & 0 & \boxed{0} \\ 0 & \boxed{0} & 1 & \boxed{0} \\ 0 & \boxed{0} & 0 & \boxed{1} \\ 0 & \boxed{0} & 0 & \boxed{0} \end{bmatrix} \quad \begin{bmatrix} 0 & \boxed{0} & 0 & \boxed{0} \\ \boxed{0} & \boxed{0} & \boxed{1} & \boxed{0} \\ 0 & \boxed{0} & 0 & \boxed{1} \\ \boxed{0} & \boxed{0} & \boxed{0} & \boxed{0} \end{bmatrix}$$

Induced Incident forward Incident backward Incident reciprocal

T-SBS To keep focus, let *incident OP* stand for incident forward observation procedure from now on. Note that in undirected graphs, incident is the same as incident reciprocal. Given an initial *seed* sample s_0 from U, let $s_1 = \alpha(s_0) \setminus s_0$ be the 1st-wave seed sample. Repeat the incident OP for s_1, which may or may not result in a non-empty 2nd-wave seed sample $s_2 = \alpha(s_1) \setminus (s_0 \cup s_1)$. Carry on this way yields the seed samples s_3, ..., s_T. The *seed sample of T-SBS* is given by $s = \bigcup_{r=0}^{T-1} s_r$.

The reference set s_{ref} of T-SBS is $s \times U$ in directed graphs or $s \times U \cup U \times s$ in undirected graphs. A motif $[M]$ is observed in the sample graph G_s iff $M \times M \in s_{\text{ref}}$. Frank and Snijders (1994) consider 1-SBS for node (1st-order) graph totals. Zhang and Patone (2017) develop the HTE for finite-order graph totals under T-SBS. The basis of inference is the graph sample inclusion probabilities.

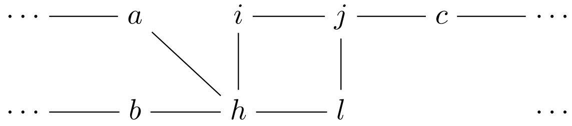


Figure 8: An example of 4-cycle $[h, i, j, l]$

However, not all the motifs observed in G_s are eligible for estimation. Consider the 4-cycle motif $[M]$ with $M = \{h, i, j, l\}$ in Figure 8. We need two waves to observe a 4-cycle, starting from any node in M, so that it is observed under 3-SBS starting from any of a, b, c. If only a is selected in s_0, then we would observe this 4-cycle under 3-SBS, as well as b as another of its ancestors, but not c, so that its sample inclusion probability under 3-SBS cannot be calculated and it remains ineligible.

Zhang and Oguz-Alper (2020) develop the theory for eligible sample motifs under T-SBS. One could carry on SBS further, until one has obtained all the ancestors of the sample motifs observed under T-SBS. One could use only the eligible sample motifs under T-SBS, in a manner resembling the modified HTE that excludes the edge notes under ACS. As explained earlier, it is also possible to modify the sampling to allow for the unmodified HTE. Zhang and Oguz-Alper (2020) develop feasible BIGS representation for T-SBS with given T. One can then apply the IWE instead of only the HTE using this BIGS-IWE strategy, tailored to the number of waves when sampling is terminated.

Targeted random walk One can envisage a discrete-time walk in a graph as travelling from one city to another via the existing roads that connect the cities. In a *random walk*, one takes randomly one of the possible roads out of the current city, repeat the same at the next city, and so on. A walk reaches gradually its *equilibrium*, if the chance that one visits a given city at a given time depends less and less on the particular starting point. The *stationary probability* that a walk takes one to a given city is the fraction of times the city is visited when the walk is at equilibrium.

Random walk in large and possibly dynamic graphs has been used in many disciplines (Masuda et al., 2017), including Google PageRank (Brin and Page, 1998), especially if the walk is fast-moving. For a connected undirected graph, the stationary probabilities of a random walk, denoted by π_i for $i \in U$, are known up to a proportional constant for the nodes visited, but not the ones yet to be visited. Thompson (2006a) applies the Metropolis-Hastings acceptance mechanism to the proposed moves, in order to achieve other targeted stationary probabilities, such as $\pi_i \propto$ degree+1. This requires one to observe *all* the neighbours of *all* the adjacent nodes of the current one, which may be impractical. Avrachenkov et al. (2010) devise an elegant random walk that requires only the knowledge of the adjacent nodes at each step time, for which the stationary probabilities at equilibrium is known up to a proportional constant for undirected graphs. The disconnected components are accommodated by random jumps via an imaginary node. We refer to it as *targeted random walk (TRW)*.

The random-walk inclusion probability of a node is intractable. Insofar as the stationary probability π_i is the same for a given node i at any time step for a random walk at equilibrium, and π_i is only known up to a proportional constant, approximately unbiased estimation is possible, e.g. for the ratio between two *node* totals using the generalised ratio estimator (Thompson, 2006a).

For other finite-order graph totals generally, Zhang (2020a) demonstrates that inference can be based on the *stationary successive sampling probability (S3P)* of any subsequence from the TRW states $\{X_0, X_1, ..., X_T\}$. For example, suppose $(X_t, X_{t+1}) = (i, j)$ where $a_{ij} = 1$. As long as both X_t and X_{t+1} belong to the seed sample s of the TRW, following the same definition of seed sample of T-SBS above, one can observe e.g. all the triangles (i, j, h) in the graph. The S3P of the actual sampling sequence $(X_t, X_{t+1}) = (i, j)$ is $\pi_i p_{ij}$, where p_{ij} is the corresponding transition probability. Moreover, all the possible (X_t, X_{t+1}) that lead to the observation of the same triangle (i, j, h) are called the *equivalent successive sampling sequences (ES3)*, including $(X_t, X_{t+1}) = (j, i), (h, i), (i, h), (h, j), (j, h)$ in addition to (i, j). The ES3 of a motif $[M]$ constitutes its multiplicity under TRW sampling sequence-by-sequence. A motif $[M]$ is eligible for estimation, if its ES3 is observed under the TRW. This yields a BIGS representation of TRW, with the motifs of interest in Ω and their ES3s in F. By this development, BIGS-IWE becomes a feasible strategy for any function of finte-order graph totals under TRW, as long as the function is invariant towards the unknown proportional constant in π_i.

5 Some topics for future research

Graph sampling is clearly the future of sampling. We have provided a brief introduction to it mainly by examples. The references contain many details that may be helpful for further reading.

A topic for future research is other possible bases of inference in various graph sampling situations. For instance, Thompson (2006b) considers adaptive web sampling where, at each wave, a subset of the already sampled nodes are used as the seeds for random walks from them. The wave-by-wave conditional sampling probabilities are used for estimation of node totals, together with all the seed sample selection probabilities. A theory is needed for other finite-order graph totals.

It is intriguing to consider other parameters than graph totals (2) and functions of them. Newman (2010) is an excellent source of candidates, many of which can be characterised as 'a local parameter dependent on the whole graph'. For example, the in-degree of a node i is a local parameter that only depends on its in-edges, but not the rest of the graph. The betweenness of a node provides an example of what we have in mind. As can be seen below, the shortest-path (SP) betweenness of ★ is 0, which is defined as the fraction of SPs between pairs of nodes in a graph passing through it, because it is always 'short-circuited' by the two ○ nodes. Whereas ★ has a high random-walk (RW) betweenness (Newman, 2005), which is defined as the fraction of RWs between pairs of nodes in a graph passing through it. How would a sampling strategy look like for such parameters?

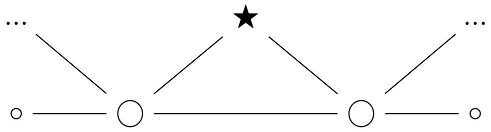


Figure 9: An example for betweenness

Graph sampling poses not least an enormous opportunity for computation. Efficient sample motif counting algorithms are obviously critical in applications, and their availability can influence the theory in return. For instance, Frank (1971) considers 'telepathy-like' observation of whether two nodes are connected, without explicating any path between them. One can envisage the possibility the relevant algorithm is so fast that it is virtually instant when the graph is known, e.g. depending on the data structure implemented. However, the graph may be so large or dynamic that sampling is still needed for 'graph compression'. The availability of such *remote* observation procedures could easily lead to other possibilities of sample graph, basis of inference and graph sampling strategy.

References

Avrachenkov, K., Ribeiro, B. and Towsley, D. (2010). Improving Random Walk Estimation Accuracy with Uniform Restarts. *Research report*, RR-7394, INRIA. inria-00520350

Becker, E.F. (1991). A terrestrial furbearer estimator based on probability sampling. *The Journal of Wildlife Management*, 55:730–737.

Birnbaum, Z.W. and Sirken, M.G. (1965). *Design of Sample Surveys to Estimate the Prevalence of lRareDiseases: Three Unbiased Estimates*. Vital and Health Statistics, Ser. 2, No.11. Washington:Government Printing Office.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30: 107–117.

Frank, O. (1971). *Statistical inference in graphs*. Stockholm: Försvarets forskningsanstalt.

Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1(3):235–264.

Frank, O. (1978). Estimation of the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5:177–188.

Frank, O. (1979). Sampling and estimation in large social networks. *Social networks*, 1(1):91–101.

Frank, O. (1980). Sampling and inference in a population graph. *International Statistical Review/Revue Internationale de Statistique*, 48:33–41.

Frank, O. (1981). A survey of statistical methods for graph analysis. *Sociological methodology*, 12:110–155.

Frank, O. (2011). Survey sampling in networks. *The SAGE Handbook of Social Network Analysis*, pages 389–403.

Frank O. and Snijders T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10:53–53.

Goodman, L.A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170.

Lavalleè, P. (2007). *Indirect Sampling*. Springer.

Masuda, N., Porter, M.A. and Lambiotte, R. (2017) Random walks and diffusion on networks. *Physics Reports*, 716-717: 1–58. http://dx.doi.org/10.1016/j.physrep.2017.07.007

Newman, M.E.J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27, 39–54.

Newman, M.E.J. (2010). *Networks: An Introduction*. Oxford University Press.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Patone, M. and Zhang, L.-C. (2020). Incidence weighting estimation under bipartite incidence graph sampling. https://arxiv.org/abs/2004.04257

Sirken, M.G. (2005). *Network Sampling*. In *Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd. DOI: 10.1002/0470011815.b2a16043

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65:257–266.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85:1050–1059.

Thompson, S.K. (1991). Adaptive cluster sampling: Designs with primary and secondary units. *Biometrics*, 47:1103–1115.

Thompson, S.K. (2006a). Targeted random walk designs. *Survey Methodology*, 32, 11–24.

Thompson, S.K. (2006b). Adaptive Web Sampling. *Biometrics*, 62, 1224–1234.

Thompson, S.K. (2012). *Sampling*. John Wiley & Sons, Inc.

Zhang, L.-C. (2020a). Targeted random walk sampling from large dynamic graphs. *Talk presented at University of Perugia, November 3, 2020.*

Zhang, L.-C. (2020b). Sampling designs for epidemic prevalence estimation. https://arxiv.org/abs/2011.08669

Zhang, L.-C. and Patone, M. (2017). Graph sampling. *Metron*, **75**, 277-299.

Zhang, L.-C. and Oguz-Alper, M. (2020). Bipartite incidence graph sampling. https://arxiv.org/abs/2003.09467

Country Reports

ARGENTINA

Reporting: **Verónica Beritich**

INDEC begins the 2020/2021 National Economic Census

Fifteen years after the last economic census, the National Institute of Statistics and Censuses (INDEC) reports that on November 30, 2020 the 2020/2021 National Economic Census begins. This statistical operation will allow knowing the updated economic structure of Argentina. Along with the population census, the economic census is one of the fundamental pillars of the country's statistical infrastructure.

The CNE 2020/2021 is an exhaustive statistical survey on people and companies that have registered activity in the Argentine Republic. It includes financial and non-financial companies, freelancers, and institutions without profit.

INDEC is planning this economic census in two stages. The first stage consists of recording all economic units with activity in the national territory in a digital statistical register. Next, the second stage will consist of carrying out several economic structural surveys and sectoral surveys to obtain information on production and inputs disaggregated by activity and product, distribution channels and margins, among other data related to non-financial corporations.

The first stage will be carried out by means of a digital questionnaire developed at INDEC, called e-CNE. More than 5,500,000 human and legal persons will complete it through the online application. This economic census aims to leave behind the classic paper questionnaires to obtain high-quality information in less time, at a lower cost compared to a territorial sweep.

Legal persons can complete the e-CNE between November 30 and January 31, 2021, according to the last number of their Unique Tax Identification Code (CUIT). With the same criteria, human persons will be able to complete it between February and June 2021. Those who participate in each specific instance will receive the summons to complete the e-CNE at their electronic fiscal address. In addition, they will have access to different tutorial videos to learn about how to enter the system and upload the information quickly and safely.

To enter to the e-CNE application, the validation of the identity of the persons will be previously required by means of their CUIT number and tax code in the Autenticar platform. Then, the system will redirect to a secure INDEC site to complete the electronic questionnaire.

Depending on the type of person, the questionnaire may have 14, 15 or 17 questions.

What questions will be answered at the e-CNE?

1. Name or business name of the company
2. CUIT
3. Location
4. Do you have establishments in more than one province?
5. Description of the main activity
6. Goods or services resulting from the main activity
7. Raw materials, materials or direct expenses for the production of goods or services
8. Annual turnover (2019)

9. Number of workers in a dependency relationship (salaried) on December 31, 2019
10. Auxiliary support units to carry out the main activity
11. Do you carry out other activities?
12. Description of secondary activities (if applicable)
13. Do you produce any good or service for your own consumption?
14. Do you produce fixed assets for your own use? (buildings, structures, machinery and equipment, computer developments, etc.)
15. Do you carry out research and development activities?

These results will allow the analysis and design of productive, sectoral and occupational public policies. They will benefit the same companies participating in the census for taking decisions in the private sector, whether at a business, academic or professional level. Likewise, they will also locate territorially the markets in terms of production and employment. They will also facilitate comparisons between the individual company and the total branch of activity or geographical area in which it is located.

With the information obtained from the e-CNE, the statistical register of economic units (REUE) will be produced which will be updated periodically with administrative records. It will allow the construction of a master sampling frame for the necessary surveys to obtain the short-term economic indicators as well as to make the change of the base year for the System of National Accounts and to build a new input-output matrix (MIP).

General information can be found at www.indec.gob.ar.

For further information, please contact ces@indec.gob.ar.

AUSTRALIA

Reporting: **Mary-Anne Stewart**

Data visualisation: understanding and conveying information more effectively

"The greatest value of a picture is when it forces us to notice what we never expected to see."

John Tukey (1915-2000)

For his biography, please see https://mathshistory.st-andrews.ac.uk/Biographies/Tukey/.

Data visualisation techniques like graphs and diagrams have been used to help us understand data and information for centuries. Good visualisation reveals stories, patterns, clusters, relationships, and anomalies like outliers or missing data, that are not as obvious or are harder to see in table and text format.

In 2020, the COVID-19 pandemic has increased the volume and style of data visualisations that we see every day in the media. Further, in recent times statisticians and data scientists are acquiring and using increasingly complex and large datasets, including administrative and transactions data. These two factors have influenced our data visualisation work program.

Data visualisation is often separated into exploratory and explanatory visualisation. Explanatory work focusses on telling stories with data, conveying information to the audience in clear and appealing ways. In contrast, exploratory visualisation is about making sense of new data sources, understanding what's going on in the data and determining how they could potentially be used in different ways. This assists analysts to assess the content of a new data source, see what's interesting and plan more directed analysis. In some situations, there is overlap between these two types.

Examples of recent data visualisation work in the Australian Bureau of Statistics (ABS) include:

- exploratory visualisation of a range of new data sources
- designing new visualisations for Merchandise Trade data processing and some specialised reports
- establishing visualisation guidelines for different types of published materials
- exploring visual techniques for interpreting results from complex "black box" analysis methods and models so they're easier to understand and explain (such as complicated machine learning methods)
- creating new types of dynamic graphs for social media. Some dynamic line plots were released in August-September 2020 showing how single touch payroll jobs have changed during the COVID-19 pandemic

Our upcoming work will include:

- exploratory visualisation of a range of new data sources and statistical methods - ABS is always investigating new sources and methods with the aim of reducing burden on survey respondents and providing richer information on the Australian economy and society
- exploring more visual techniques to support effective statistical processing - for example allowing quicker identification of anomalies and confirmation of clean data ready for release
- building ABS capability in different types of data visualisation

For more information, please contact Mary-Anne Stewart at methodology@abs.gov.au.

CANADA

Reporting: **Kenneth Chu**

Use of Machine Learning Techniques for Crop Yield Prediction

Statistics Canada recently completed a research project for the Field Crop Reporting Series (FCRS) [1] on the use of supervised machine learning techniques for early-season crop yield prediction, for the Canadian province of Manitoba, based on local vegetation remote sensing data and weather measurements from January to July.

A number of prediction techniques were examined, including random forests, support vector machines, elastic-net regularized generalized linear models, and multilayer perceptrons. Accuracy and computation time considerations led us to focus attention on XGBoost [2] with linear base learner.

The main contribution of the research project is the adaptation of **rolling window forward validation** (RWFV) [3] as hyperparameter tuning protocol. RWFV is a special case of forward validation [3], a family of validation protocols designed to prevent temporal information leakage for supervised learning based on time series data.

In order to evaluate the performance of this prediction strategy based on XGBoost (Linear) and RWFV, we computed the series of prediction errors that would have resulted had the strategy actually been deployed for past production cycles. In other words, these prediction errors of virtual past production cycles were regarded as estimates of the generalization error within the statistical production context of the FCRS.

The resulting prediction strategy has exhibited smaller prediction errors than the method currently deployed in production (variable selection via Lasso, followed by robust linear regression), consistently over consecutive historical production runs. This strategy has been implemented in a production-ready R package. The strategy has entered the final pre-production testing phase, to be jointly conducted by subject matter experts and the agricultural programme methodologists.

The results of this research project were presented in October 2020 during the Machine Learning Virtual Sessions, organized by the High-Level Group for the Modernisation of Official Statistics, United Nations Economic Commission for Europe [4].

For more information, please contact kenneth.chu@canada.ca.

References and further information

[1] https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3401

[2] Chen, Tianqi; Guestrin, Carlos (2016). *XGBoost: A Scalable Tree Boosting System*. In Krishnapuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev (eds.). Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM. pp. 785–794. https://arxiv.org/abs/1603.02754. https://dl.acm.org/doi/10.1145/2939672.2939785.

[3] Schnaubelt, Matthias (2019). *A comparison of machine learning model validation schemes for non-stationary time series data*, FAU Discussion Papers in Economics, No. 11/2019, Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Economics, Nürnberg. https://www.econstor.eu/handle/10419/209136

[4] Bosa, Keven; Chu, Kenneth (2020). *Deploying Machine Learning Techniques for Crop Yield Prediction*. HLG-MOS Machine Learning Project, High-Level Group for the Modernisation of Official Statistics, United Nations Economic Commission for Europe. https://statswiki.unece.org/display/ML/Other+applications+of+Machine+Learning

FIJI

Reporting: **M.G.M. Khan**

Recent and Upcoming Surveys at the Fiji Bureau of Statistics

The Fiji Bureau of Statistics (FBoS) has completed the 2019-2020 *Household Income and Expenditure Survey (HIES)* and a *post-survey evaluation* is being conducted for the very first time. Surveys on *2nd Demand Side Survey of financial products* in partnership with Reserve Bank of Fiji (central bank) and *Multiple Indicator Survey (MICS)* with UNICEF will also begin. FBoS is also preparing for other two ad-hoc surveys which are "*Impact Evaluation of Secured Transaction Reform in Fiji*" in partnership with Reserve Bank of Fiji and *Office on Drugs and Crime (UNODC) regarding human trafficking* in conjunction with United Nations.

Apart from 2019-2020 HIES, all the other surveys mentioned above are being carried out for the very first time.

NEW ZEALAND

Reporting: **Mark Turner** and **James McKay**

COVID-19 Data Portal

COVID-19 has accelerated the pace of decision-making, triggering a need for more near-real-time economic, social and health data in one place.

Adopting a customer-centric Agile approach, Stats NZ proactively engaged with key government and private sector clients to understand and meet their data needs as quickly and as efficiently as possible. Stats NZ found that the majority of clients, while appreciating the accuracy and coherence of Stats NZ's traditional data products (e.g., GDP), were prepared to sacrifice these aspects for better accessibility to timelier and more frequent data on COVID-19.

Stats NZ's COVID-19 Data Portal is a collection of primarily non-official indicators relevant for understanding the economic and social recovery from the impacts of COVID-19 through a wellbeing perspective. Given the content of the portal it is hosted in the experimental section of the Stats NZ website, so as to distinguish the indicators from our official statistics.

The COVID-19 Data Portal was built to be a generic data hosting framework. It is flexible enough to consume any machine-readable dataset and allows us to define custom visualisations of that data. This is achieved by making clear application programming interfaces between the data ingestion, a consistent data and metadata structure, and the visualisations.

We are able to automate the ingestion of a range of data formats by having a flexible ingestion layer. The framework has a defined place for small custom functions to handle non-standard input data, which interface with the consistent data format used through the rest of the program. This means we have been able to quickly bring in new data sets with minimal burden on the data suppliers. Because this process is fully automated, with a small amount of up-front effort, there is no on-going burden on Stats NZ for maintaining and updating this data in the future. As a result, this framework has enabled us to scale the number of data sets massively, with no significant increase in the time required to keep it up to date.

This new, highly relevant, data product was relatively inexpensive to develop, and to a large extent leveraged existing data holdings and/or capabilities. Stats NZ is ready to coordinate the expansion of this product as required or terminate indicators as and when they are no longer needed.

COVID 19 Data Portal: https://www.stats.govt.nz/experimental/covid-19-data-portal

Open-source code is available: https://github.com/StatisticsNZ/data_portal

For further information, please contact james.mckay@stats.govt.nz

Upcoming Conferences and Workshops

ICES VI - The International Conference on Establishment Statistics will take place 14-17 June 2021 in New Orleans, U.S. Website: https://ww2.amstat.org/meetings/ices/2020/conferenceinfo.cfm/

SAE2021 - BigSmall - Conference on Small Area Estimation, with the theme "Big Data for Small Areas", will be held 20-25 September in Naples, Italy, as a satellite conference to the World Statistics Conference in 2021. Website: https://sae2020.org/

11e Colloque International Francophone sur les Sondages - 11th International Francophone Conference on Surveys will take place 6-8 October 2021 in Brussels, Belgium. Website: http://sondages2020.sciencesconf.org

Other Conferences on survey statistics and related areas

American Statistical Association Conference on Statistical Practice is planned to take place 18-20 February 2021 in Nashville, USA.

Website: https://ww2.amstat.org/meetings/csp/2020/index.cfm

Statistics in the Big Data Era will take place 2-4 June 2021 in Berkeley, USA. Website: https://simons.berkeley.edu/workshops/statistics-big-data-era

Symposium on Data Science & Statistics is planned to take place June 2-5, 2021 in Missouri, USA. Website: https://ww2.amstat.org/meetings/sdss/2021/

ANZSC 2021 – Australian Statistical Society and New Zealand Statistical Association Conference will take place 5-9 July 5, Gold Coast, Australia. Website: https://anzsc2021.com.au/

Conference and Special Issue of Journal of the Royal Statistical Society Series A in memory of Fred Smith and Chris Skinner will be held in Southampton, UK, 8-10 July 2021. See the August 2020 IASS Newsletter for full details.

63rd ISI World Statistics Congress will take place 11-15 July 2021 and will be virtual. Website: https://www.isi2021.org/

Joint Statistical Meetings 2021 will take place 7-12 August in Seattle, USA. Website: https://www.amstat.org/ASA/Meetings/Joint-Statistical-Meetings.aspx

2021 Women in Statistics and Data Science Conference will take place 7-9 October in Pittsburgh, USA. Website: https://ww2.amstat.org/meetings/wsds/2021/

The Survey Research Methods Section (SRMS) of the ASA Information on activities of the Survey Research Methods Section of the American Statistical Association (ASA) are available at: https://community.amstat.org/surveyresearchmethodssection/home

International Statistical Institute calendar of events is at https://www.isi-web.org/events-and-awards/calendar

In Other Journals

Journal of Survey Statistics and Methodology

Volume 8, Issue 4, September 2020

https://academic.oup.com/jssam/issue/8/4

Survey Methodology

Do Forced-Choice (FC) Questions Trigger Deeper Cognition than Check-All-That-Apply (CATA) Questions?
Cornelia E Neuert

The Effect of Question Characteristics on Question Reading Behaviors in Telephone Surveys
Kristen Olson, Jolene D Smyth, Antje Kirchner

So Many Questions, So Little Time: Integrating Adaptive Inventories into Public Opinion Research
Jacob M Montgomery, Erin L Rossiter

On Examining the Quality of Spanish Translation in Telephone Surveys: A Novel Test-Retest Approach
Robert P Agans, Quirina M Vallejos, Thad S Benefield

Is That Still the Same? Has that Changed? On the Accuracy of Measuring Change with Dependent Interviewing
Annette Jäckle, Stephanie Eckman

Interventions On-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates
Stephanie Coffey, Benjamin Reist, Peter V Miller

Split-Sample Design with Parallel Protocols to Reduce Cost and Nonresponse Bias in Surveys
Andy Peytchev

Survey Statistics

A Permutation Test on Complex Sample Data
Daniell Toth

Bayesian Inference of Finite Population Quantiles for Skewed Survey Data Using Skew-Normal Penalized Spline Regression
Yutao Liu, Qixuan Chen

Volume 8, Issue 5, November 2020

https://academic.oup.com/jssam/issue/8/5

Survey Methodology

Improving Survey Response Rates with Visible Money
Matthew Debell, Natalya Maisel, Brad Edwards, Michelle Amsbary, Vanessa Meldener

The Impact of Varying Financial Incentives on Data Quality in Web Panel Surveys
Thomas Luke Spreen, Lisa A House, Zhifeng Gao

What They Expect Is What You Get: The Role of Interviewer Expectations in Nonresponse to Income and Asset Questions
Sabine Friedel

Estimation of Underreporting in Diary Surveys: An Application Using the National Household Food Acquisition and Purchase Survey
Mengyao Hu, John A Kirlin, Brady T West, Wenyi He, Ai Rene Ong, Shiyu Zhang, Xingyou Zhang

Who Can You Count On? Understanding The Determinants of Reliability
Roger Tourangeau, Ting Yan, Hanyu Sun

Survey Statistics

Measures of the Degree of Departure from Ignorable Sample Selection
Roderick J A Little, Brady T West, Philip S Boonstra, Jingwei Hu

Multiple Imputation with Survey Weights: A Multilevel Approach
M Quartagno, J R Carpenter, H Goldstein

Applications

Accuracy in the Application of Statistical Matching Methods for Continuous Variables Using Auxiliary Data
Arnout Van Delden, Bart J Du Chatinier, Sander Scholtus

Survey Methodology, June 2020, Vol. 46, no. 1

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2020001-eng.htm

Are probability surveys bound to disappear for the production of official statistics?
Jean-François Beaumont

Local polynomial estimation for a small area mean under informative sampling
Marius Stefan and Michael A. Hidiroglou

Small area estimation methods under cut-off sampling
María Guadarrama, Isabel Molina and Yves Tillé

Model-assisted sample design is minimax for model-based prediction
Robert Graham Clark

Considering interviewer and design effects when planning sample sizes
Stefan Zins and Jan Pablo Burgard

A new double hot-deck imputation method for missing values under boundary conditions
Yousung Park and Tae Yeon Kwon

Survey Methodology, December 2020, Vol. 46, no. 2

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2020002-eng.htm

Estimation and inference of domain means subject to qualitative constraints
Cristian Oliva-Aviles, Mary C. Meyer and Jean D. Opsomer

Bayesian hierarchical weighting adjustment and survey inference
Yajuan Si, Rob Trangucci, Jonah Sol Gabry and Andrew Gelman

Firth's penalized likelihood for proportional hazards regressions for complex surveys
Pushpal K. Mukhopadhyay

Probability-proportional-to-size ranked-set sampling from stratified populations
Omer Ozturk

Semi-automated classification for multi-label open-ended questions
Hyukjun Gweon, Matthias Schonlau and Marika Wenemark

Journal of Official Statistics

JOS
Journal of Official Statistics
September 2013

Volume 36 (2020): Issue 3 (Sep 2020): Special Issue on Nonresponse

https://content.sciendo.com/view/journals/jos/36/3/jos.36.issue-3.xml

Preface
Edith de Leeuw, Annemieke Luiten, and Ineke Stoop

Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys
Annemieke Luiten, Joop Hox, and Edith de Leeuw

Continuing to Explore the Relation between Economic and Political Factors and Government Survey Refusal Rates: 1960–2015
Luke J. Larsen, Joanna Fane Lineback, and Benjamin M. Reist

Evolution of the Initially Recruited SHARE Panel Sample Over the First Six Waves
Sabine Friedel and Tim Birkenbach

The Action Structure of Recruitment Calls and Its Analytic Implications: The Case of Disfluencies
Bo Hee Min, Nora Cate Schaeffer, Dana Garbarski, and Jennifer Dykema

Measurement of Interviewer Workload within the Survey and an Exploration of Workload Effects on Interviewers' Field Efforts and Performance
Celine Wuyts and Geert Loosveldt

Assessing Interviewer Performance in Approaching Reissued Initial Nonrespondents
Laurie Peeters, David De Coninck, Celine Wuyts, and Geert Loosveldt

Implementing Adaptive Survey Design with an Application to the Dutch Health Survey
Kees van Berkel, Suzanne van der Doef, and Barry Schouten

The Effects of Nonresponse and Sampling Omissions on Estimates on Various Topics in Federal Surveys: Telephone and IVR Surveys of Address-Based Samples
Floyd J. Fowler, Philip Brenner, Anthony M. Roman, and J. Lee Hargraves

Working with Response Probabilities
Jelke Bethlehem

A Validation of R-Indicators as a Measure of the Risk of Bias using Data from a Nonresponse Follow-Up Survey
Caroline Roberts, Caroline Vandenplas, and Jessica M.E. Herzing

Proxy Pattern-Mixture Analysis for a Binary Variable Subject to Nonresponse
Rebecca R. Andridge and Roderick J.A. Little

Volume 36 (2020): Issue 4 (Dec 2020)

https://content.sciendo.com/view/journals/jos/36/4/jos.36.issue-4.xml

Letter to the Editors
Andreas V Georgiou

Basic Statistics of Jevons and Carli Indices under the GBM Price Model
Jacek Białek

Developing Land and Structure Price Indices for Ottawa Condominium Apartments
Kate Burnett-Isaacs, Ning Huang, and W. Erwin Diewert

An Improved Fellegi-Sunter Framework for Probabilistic Record Linkage Between Large Data Sets
Marco Fortini

Three-Form Split Questionnaire Design for Panel Surveys
Paul M. Imbriano and Trivellore E. Raghunathan

Double Barreled Questions: An Analysis of the Similarity of Elements and Effects on Measurement Quality
Natalja Menold

The Representativeness of Online Time Use Surveys. Effects of Individual Time Use Patterns and Survey Design on the Timing of Survey Dropout
Petrus te Braak, Joeri Minnen, and Ignace Glorieux

Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context
James Wagner, Brady T. West, Michael R. Elliott, and Stephanie Coffey

Book Review: Paul C. Beatty, Debbie Collins, Lyn Kaye, Jose-Luis Padilla, Gordon B. Willis, and Amanda Wilmot. Advances in Questionnaire Design, Development, Evaluation and Testing. 2019, Wiley, ISBN: 978-1-119-26362-3, 816 pages
Jennifer Edgar

Book Review: Yuling Pan, Mandy Sha, and Hyunjoo Park. The Sociolinguistics of Survey Translation. 2020, New York: Routledge, ISBN 978-1-138-55087-2, 166 pages
Patricia Goerman

Survey Practice

Vol. 13, Issue 1, 2020

https://www.surveypractice.org/issue/1854

Articles

Moving Beyond Sex: Measuring Gender Identity in Telephone Surveys
Dan Cassino

Minnesota Social Contacts and Mixing Patterns Survey with Implications for Modelling of Infectious Disease Transmission and Control
Audrey M. Dorélien, Alisha Simon, Sarah Hagge, Kathleen Thiede Call, Eva Enns, Shalini Kulasingam

Interviewer Face Coverings and Response to Personal Visit Surveys: A Case Study of the 2020 U.S. Census
Nancy Bates, Laura Kail, Amanda Price

Typographic Cueing Facilitates Survey Completion on Smartphones in Older Adults
Brian Falcone, Christopher Antoun, Elizabeth Nichols, Erica Olmsted-Hawala, Ivonne Figueroa, Alda Rivas, Shelley Feuer, Lin Wang

Adapting surveys to the modern world: Comparing a research messenger design to a regular responsive design for online surveys
Vera Toepoel, Peter Lugtig, Bella Struminskaya, Anne Elevelt, Marieke Haan

Scale-sensitive response behavior!? Consequences of offering versus omitting a “don’t know” option and/or a middle category
Daniela Wetzelhütter

Design Considerations for Live Video Survey Interviews
Michael F. Schober, Frederick G. Conrad, Andrew L. Hupp, Kallan M. Larsen, Ai Rene Ong, Brady T. West

The Effect of Incentives and Mode of Contact on the Recruitment of Teachers into Survey Panels
Michael Robbins, Jennifer Hawes-Dawson

Collecting Online Survey Data: A Comparison of Data Quality among a Commercial Panel & MTurk
Bingbing Zhang, Sherice Gearhart

Survey Research Methods

Journal of the European Survey Research Association

Vol 14 No 3 (2020)

https://ojs.ub.uni-konstanz.de/srm/issue/view/222

Designing Better Questions for Complex Concepts with Reflective Indicators
Willem Saris, Irmtraud Gallhofer

Automatic Coding of Open-ended Questions into Multiple Classes: Whether and How to Use Double Coded Data
Zhoushanyue He, Matthias Schonlau

Response Patterns in a Multi-day Diary Survey: Implications for Adaptive Survey Design
Mengyao Hu, Edmundo Roberto Melipillán, Brady T. West, John A. Kirlin, Ilse Paniagua

The More Similar, the Better? How (Mis)Match Between Respondents and Interviewers Affects Item Nonresponse and Data Quality in Survey Situations
Felix Bittmann

Memory Effects in Repeated Survey Questions: Reviving the Empirical Investigation of the Independent Measurements Assumption
Hannah Schwarz, Melanie Revilla, Wiebke Weber

Vol 14 No 4 (2020): Special Issue: Measurement Equivalence: Testing for It and Explaining Why It is Absent

https://ojs.ub.uni-konstanz.de/srm/issue/view/218

Measurement Invariance: Testing for It and Explaining Why It is Absent
Katharina Meitinger, Eldad Davidov, Peter Schmidt, Michael Braun

The Comparability of Measures in the Ageism Module of the Fourth Round of the European Social Survey, 2008-2009
Daniel Seddig, Dina Maskileyson, Eldad Davidov

Revisiting the Middleton Alienation Scale: In Search of a Cross-Culturally Valid Instrument
Ekaterina Lytkina

Investigating the Relative Impact of Different Sources of Measurement Non-Equivalence in Comparative Surveys An Illustration with Scale Format, Data Collection Mode and Cross-National Variations
Caroline Roberts, Oriane Sarrasin, Michèle Ernst Stähli

Measurement Equivalence of Subjective Well-Being Scales under the Presence of Acquiescent Response Style for the Racially and Ethnically Diverse Older Population in the United States
Sunghee Lee, Elizabeth Vasquez, Lindsay Ryan, Jacqui Smith

Vol 14 No 5 (2020)

https://ojs.ub.uni-konstanz.de/srm/issue/view/224

Valid vs. Invalid Straightlining: The Complex Relationship Between Straightlining and Data Quality
Kevin Reuning, Eric Plutzer

The Estimation of Voter Transitions in the 2015 British General Election: Combining Online Panels and Aggregate Data at the Constituency Level
Paul W. Thurner, Ingrid Mauerer, Maxim Bort, André Klima, Helmut Küchenhoff

Collecting biomarkers in Australian primary schools: Insights from the field
Mandy Truong, Mienah Z Sharif, Rebecca Moorhead, Jeffrey Craig, Pamela Leong, Yuk Jing Loke, Kevin Dunn, Naomi Priest

Helpful Reminders? Health Survey Participation and Doctor's Visits among Aging Adults
Jennifer Caputo

Comparing the participation of Millennials and older age cohorts in the CROss-National Online Survey panel and the German Internet Panel
Melanie Revilla, Jan K. Höhne

The Benefits of Conversational Interviewing Are Independent of Who Asks the Questions or the Types of Questions They Ask
Frost A Hubbard, Frederick G Conrad, Christopher Antoun

Other Journals

- **Statistical Journal of the IAOS**
  - https://content.iospress.com/journals/statistical-journal-of-the-iaos/
- **International Statistical Review**
  - https://onlinelibrary.wiley.com/journal/17515823
- **Transactions on Data Privacy**
  - http://www.tdp.cat/
- **Journal of the Royal Statistical Society, Series A (Statistics in Society)**
  - https://rss.onlinelibrary.wiley.com/journal/1467985x
- **Journal of the American Statistical Association**
  - https://amstat.tandfonline.com/toc/uasa20/current
- **Statistics in Transition**
  - https://sit.stat.gov.pl

Welcome New Members!

We are very pleased to welcome the following new IASS members!

Title	First name	Surname	Country
MR.	Calogero	Carletto	United States
PROF.	Jan	Van den Brakel	Germany
MRS.	Lucia	Spoiala	Moldova
MR.	Mathias Mulumba	Zungu	Uganda

IASS Executive Committee Members

Executive officers (2019 – 2021)

President:	Denise Britz do Nascimento Silva (Brazil)	denisebritz@gmail.com
President-elect:	Monica Pratesi (Italy)	monica.pratesi@unipi.it
Vice-Presidents:		
Scientific Secretary:	James Chipperfield (Australia)	james.chipperfield@abs.gov.au
VP Finance:	Lucia Barroso (Brazil)	lpbarroso@gmail.com
Chair of the Cochran-Hansen Prize Committee and IASS representative on the ISI Awards Committee:	Isabel Molina (Spain)	imolina@est-econ.uc3m.es
IASS representatives on the World Statistics Congress Scientific Programme Committee:	Cynthia Clark (USA) in 2017-2019	czfclark@cox.net
	Monica Pratesi (Italy)	monica.pratesi@unipi.it
IASS representative on the World Statistics Congress short course committee:	Nadia Lkhoulf (Morocco)	n.lkhoulf@hcp.ma
Ex Officio Member:	Ada van Krimpen	an.vankrimpen@cbs.nl

IASS Twitter Account @iass_isi (https://twitter.com/iass_isi)

Institutional Members

International organisations:

- Eurostat (European Statistical Office)
- Observatoire économique et statistique d'Afrique subsaharienne (AFRISTAT)

National statistical offices:

- Australian Bureau of Statistics, Australia
- Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil
- Statistics Canada, Canada
- Statistics Denmark, Denmark
- Statistics Finland, Finland
- Statistisches Bundesamt (Destatis), Germany
- Israel Central Bureau of Statistics, Israel
- Istituto nazionale di statistica (Istat), Italy
- Statistics Korea, Republic of Korea
- Direcção dos Serviços de Estatística e Censos (DSEC), Macao, SAR China
- Statistics Mauritius, Mauritius
- Instituto Nacional de Estadística y Geografía (INEGI), Mexico
- Statistics New Zealand, New Zealand
- Statistics Norway, Norway
- Instituto Nacional de Estatística (INE), Portugal
- Statistics Sweden, Sweden
- National Agricultural Statistics Service (NASS), United States
- National Center of Health Statistics (NCHS), United States

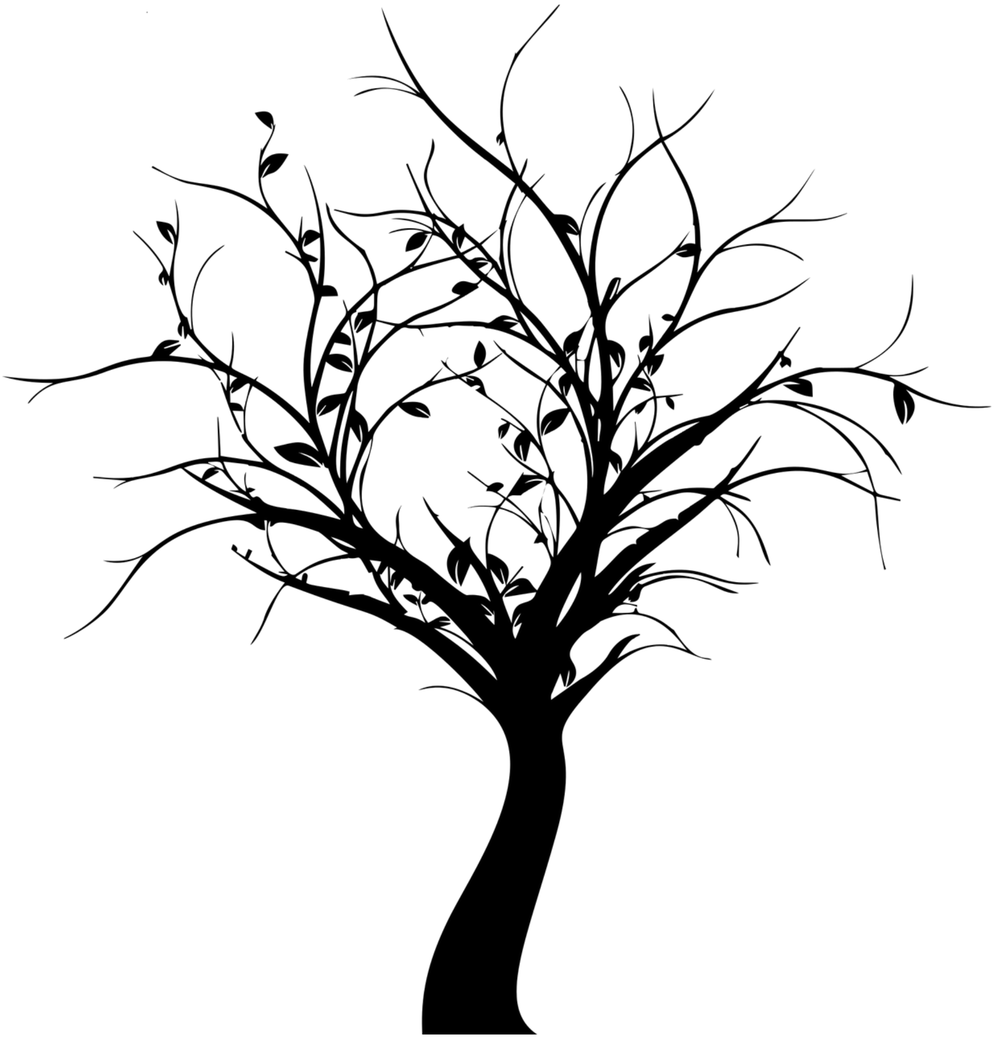
Private companies:

- Numérika (Asesoría estadística y estudios cuantitativos), Mexico
- RTI International, United States
- Survey Research Center (SRC), United States
- Westat, United States

Save a tree!

Read *the Survey Statistician* online!

http://isi-iass.org/home/services/the-survey-statistician/



Please contact Margaret de Ruiter-Molloy (m.deruitermolloy@cbs.nl)
if you would like to cancel receiving paper copies of this Newsletter.