Challenges in the Production of Official Statistics with New Methods of Data Collection

Danny Pfeffermann

SOCIETY OF STATISTICS, COMPUTER AND APPLICATIONS

December- 2023

Work as Israel's National Statistician and Director of the ICBS

I served on these two roles during **2013-2022-** a total **of 9 years**.

The **ICBS** is in charge of the **production** and **publication** of the **official statistics** in Israel.

Data are collected via **Business surveys**, **Household Surveys**, **Persons' surveys**, **administrative files** and **censuses. ≈ 65** surveys**.**

Surveys are conducted via **telephone**, **face-to-face** interviews, the **internet** and by **mail**. Often a **combination** of several collection modes.

❖ Information is published on every aspect of the life of the society.

The ICBS is a member of many international organizations**; UN**, **OECD**, **Eurostat**, **World bank**, **IMF** and the **ILO**. The organizations use the **ICBS** data for their analysis and publications.

The ICBS work during the Covid-19 pandemic

1- Our workforce reduced initially to **1/3** of the usual workforce.
Many employees had to work from home. Workforce increased later.

2- Developed with Israel’s cyber authority a protected system to permit employees to access the **ICBS** secured data from home.

3- Had to stop **face-to-face** interviews and visits to shops for price collection. Used Telephone interviews and online data instead. Used credit card purchases to complement the household expenditure survey data. This survey is used also to compute the weights for the consumer price index (**CPI**).

4- Conducted **11** special business surveys and **4** household surveys from “one day to the next.” These surveys were requested by Government ministries and the Central Bank of Israel.

5- Produced and published the **LFS** and **consumer confidence** survey estimates every two weeks, instead of every month.

6- Conducted a serological survey in a big city in Israel. Proposed and designed several other such surveys, including a national survey of children, but didn’t get the support of our Health ministry.

7- Had to use all kinds of special procedures for estimating **trends** and **seasonally adjusted** series for our survey data.

8- Developed a model for estimating the monthly **excess of mortality** resulting from the pandemic.

❖ All this work carried out with our reduced workforce, in parallel to the normal standard activity, which continued as usual.

What is official statistics? Why is it important?

Publication by a **national statistical office** (**NSO**), based on a survey, census, administrative data, and possibly **big data**.

- **Official Statistics (OS)** is what people hear of almost daily. Unemployment rates, price indexes, education attainments, poverty measures, apartment's prices, health and environmental statistics**…**
- For most people, **OS** is what statistics is all about**!!**
- **OS** is what policy makers use (**should use**) for planning and decision making.
- **Growing** demands for **detailed/timely** data, huge technological developments, declining response rates, tightened budgets**…**

⇒ **Big new challenges**

Main methods of data collection for official statistics

1- **Surveys**, based on probability samples; still the most common, and in many ways the most reliable method.

2- **Administrative records;** often requires linking several big files, which can be problematic and increase privacy concerns.

3- **Censuses**

4- **Big data;** despite all the noise, **not really implemented yet for OS;** increased pressure on **NSO's** all over the world to **digitise** (**"modernise"**) their data sources.

5- **Combinations** of the methods above.

Major problems with the use of traditional sample surveys

Yields **unbiased** estimators under the randomization (design-based) distribution without the need for statistical models. Accommodates calculation of measures of errors. **However,**

❖ Often requires **large samples** for needed level of accuracy, particularly for small domains **(SAE)**, which can be very **costly**.

❖ Sampling designs often **informative**↔ $f_s(y_i \mid i \in s, \mathrm{x}_i) \neq f_p(y_i \mid \mathrm{x}_i)$.

* Important when modelling sample data for inference on population.

❖ People and establishments are less and less willing to participate in surveys ⇒ **declining rates of response**, often **NMAR** ⇒ risk of **biased inference** if not handled properly.

Proxi surveys (one reports for many)

One person of household (**HH)** (whoever is reached), provides information for **all other members** of the household.

One of possible ways to deal with small sample sizes and nonresponse**;** very common in **HH** surveys (**e.g., LFS**, **HES**, **Censuses…**).

Possible ethical problem: Do other **HH** members agree that their **personal** data **(e.g., medical information)** is provided to interviewer**?**

❖ Major problem in **non-mandatory** surveys.

High propensity for **nonresponse: "Don't know"**.

High propensity for **correlated measurement errors**.

Example of estimates from Labor Force survey (LFS) in Israel

Estimates based on Total, Self- and Proxy respondents

Participation in Labor force & employment by gender; percentages

	Participation		Employed		Unemployed	
	Male	Female	Male	Female	Male	Female
All Sample	70.9	60.3	68.0	57.7	4.1	4.3
Self Resp.	76.3	65.3	73.3	62.8	3.9	3.9
Proxy Resp.	67.4	56.9	64.5	54.3	4.3	4.7

Different estimates obtained from **Self Resp.** and **Proxy Resp.**

Major problems with use of traditional sample surveys (cont.)

- **Timeliness-** Traditional surveys often take many months- users nowadays require that data be collected and released **"in real time"**.
- **NSOs** need to **stay relevant** in a dynamic changing world.
- **But** sometimes, survey data are much quicker than administrative files. **Example:** business **income information**.
- **Mode effects- mixed mode surveys:** different modes of response**;** **telephone**, **personal interview**, **email, internet**… different modes often offered sequentially to non-respondents with a previous mode.

Mode effects (cont.)

Mode-effects encompass two confounded effects:

Selection effect; different characteristics of respondents with different modes ⇒ **possible differences in values of target study variables**,

Measurement effect; effect of **responding differently by same person**, depending on the mode of response.

Big differences often observed in answers with different modes.

Reasons for mixed mode surveys: possible increased response, some modes **cheaper than others** (**internet!!**).

Example of mode effects- Agriculture Census, Israel, 2018

- **210** farmers responded both by internet and by telephone!!
- Ideal for assessing existence of **measurement effects**.

Study variables	**# Farmers T=I**	**# Farmers T>I**	**# Farmers T<I**
# of workers	**131**	**39**	**40**
Cultivated area	**139**	**38**	**33**

Study variables	**Mean Internet (I)**	**Mean Telephone(T)**	**Mean for T>I**	**Mean for T<I**
# of workers	**5.9**	**5.8**	**T= 15.5** **I= 7.0**	**T= 7.5** **I=17.0**
Cultivated area	**108.5**	**105.9**	**T= 318.4** **I= 192.0**	**T= 88.3** **I= 144.5**

Mode effects (cont.)

A common approach to deal with mode effects: assume that one of the modes has **no measurement effect** ⇒ by restricting to this mode, estimates of population parameters would be unbiased.

Uses **observational study** theory; requires knowledge of covariates satisfying strong ignorability conditions.

❖ No such mode guaranteed - not clear how to test its existence.

See **Pfeffermann, *JSSAM*, 2015, De Leeuw et al., Sur. Res. Methods, 2018** and **Pfeffermann & Preminger, *Sankhya*, A, 2021** for review of this and other methods.

Adjusting for mode effects-
Pfeffermann and Preminger, Sankhya, A, 2021

Consider a population ***P*** of ***N*** units and denote by (Y_i, M_i, x_i) the true outcome value, the mode used and auxiliary (covariate) values corresponding to unit $i \in P$. Suppose there are $\boldsymbol{m} = \mathbf{1,...,\overleftrightarrow{M}}$ modes with the last mode consisting of **nonrespondents**.

Assumption- for each $j \in P$ exists a **true** Y_j with ***pdf*** $f_{\boldsymbol{p}}(Y_j \mid \mathrm{x}_j)$.

❖ **Not assumed** that Y is measured accurately under **any mode**.

In what follows we consider **3 models:**

$f_{\boldsymbol{p}}(Y \mid \mathrm{x})$**;** $\Pr(M \mid Y, \mathrm{x})$**;** $f(y \mid Y, M, \mathrm{x})$**;**

$\boldsymbol{y}$- value measured for responding unit i (may **differ** from $\boldsymbol{Y}$).

Adjusting for selection effects (no measurement errors)

Assume for convenience Y=(0,1).

$$\Pr(Y_i \mid \mathrm{x}_i, M_i = m) \overset{\textit{Bayes}}{=} \frac{\Pr(M_i = m \mid Y_i, \mathrm{x}_i)\Pr_p(Y_i \mid \mathrm{x}_i)}{\Pr(M_i = m \mid \mathrm{x}_i)}.$$

$\Pr_p(Y_i \mid \mathrm{x}_i) \rightarrow$ **Target probability in the population.**

$\Pr(Y_i \mid \mathrm{x}_i, M_i = m) \rightarrow$ **Accounts for selection effects of respondents using mode m.**

❖ Requires modelling $\Pr(M_i = m \mid Y_i, \mathrm{x}_i)$ **e.g., multivariate logistic.**

Covariates explaining chosen mode not necessarily the same as covariates explaining the outcome. (For **model identification**, the two sets of covariates need to **differ** in at least one variable.)

Adjusting also for measurement effects

Denote by y_i the value measured for **responding** unit i, which in the case of measurement effects may differ from the true value Y_i.

Account for possible **measurement effects** by modelling,

$$\Pr(y_i \mid \mathrm{x}_i, M_i) = \sum_{j=0}^{1} \Pr(y_i \mid Y_i = j, \mathrm{x}_i, M_i) \frac{\Pr(M_i \mid Y_i = j, \mathrm{x}_i) \Pr_{p}(Y_i = j \mid \mathrm{x}_i)}{\Pr(M_i \mid \mathrm{x}_i)}.$$

Note: We only observe the values $\{(y_i, \mathbf{x}_i, M_i)\}$, and for the non-respondents, only the mode $\overset{\leftrightarrow}{M}$ and $\mathbf{x}$.

Model for non-respondents:

$\Pr(M_i = \overset{\leftrightarrow}{M} \mid Y_i, \mathrm{x}_i) = 1 - \sum_{m \neq \overset{\leftrightarrow}{M}} \Pr(M_i = m \mid Y_i, \mathrm{x}_i) \rightarrow$ accounts for **NMAR** nonresponse. The probability not to respond depends on Y.

Estimation of model parameters

The models defined before depend on unknown parameters. Denote, $\delta = (\alpha, \beta)'$.

$$\Pr(Y_i \mid \mathrm{x}_i, M_i; \boldsymbol{\delta}) = \frac{\Pr(M_i = m \mid Y_i, \mathrm{x}_i; \boldsymbol{\beta}) \Pr(Y_i \mid \mathrm{x}_i; \boldsymbol{\alpha})}{\Pr(M_i \mid \mathrm{x}_i; \boldsymbol{\delta})}.$$

$$\Pr(y_i \mid \mathrm{x}_i, M_i; \boldsymbol{\gamma}, \boldsymbol{\delta}) = \sum_{j=0}^{1} \Pr(y_i \mid Y_i = j, \mathrm{x}_i, M_i; \boldsymbol{\gamma}) \frac{\Pr(M_i \mid Y_i = j, \mathrm{x}_i; \boldsymbol{\beta}) \Pr(Y_i = j \mid \mathrm{x}_i; \boldsymbol{\alpha})}{\Pr(M_i \mid \mathrm{x}_i; \boldsymbol{\delta})}$$

❖ See the article for the **likelihood equations** under the models, with and without measurement effects, maximization procedures and asymptotic properties of the **MLEs**.

Prediction of finite population means

Replacing the unknown model parameters by their **MLE** permits estimating the population mean (**proportion**) $\bar{Y}_{(P)} = \frac{1}{N}\sum_{j=1}^{N} Y_j$.

Let $\hat{\rho}_i = \hat{\Pr}(Y_i = 1 \mid \mathrm{x}_i)$.

When the covariates $\{\mathrm{x}_j\}$ are known for all the population units,

$\hat{\bar{Y}}_{\text{model}} = \frac{1}{N}\sum_{j=1}^{N} \hat{\rho}_i$.

When the covariates are only known for the sampled units,

$\hat{\bar{Y}}_{\text{Hajek,model}} = \sum_{i=1}^{n} \pi_i^{-1}\hat{\rho}_i / \sum_{i=1}^{n} \pi_i^{-1}$. $\boldsymbol{\pi_i = Pr(i \in s)}$.

- See the article for results of a **simulation study**, including the estimation of the means **for each mode** and testing of the model.

Application to 2017 Crime Victimization Survey in Israel

Probability sample, collects data on victimization and socio-demographic characteristics. Total sample size $n = 7035$, with **11%** responding via the internet, **60%** by telephone, and **29%** not responding. (**3 modes**.)

❖ **41.4%** of the internet respondents and **23.5%** of the telephone respondents have an academic degree, suggesting the existence of mode **selection effects**, and possibly also measurement effects.

❖ The target variable of interest is the binary variable of **having an academic degree**. The **ICBS** has an extensive register of education, with coverage of over than **95%**, allowing to compare our predictors of the population proportion with the **truth**, $\bar{Y}_{(P)} = 0.24$.

Models fitted

Logistic models with the following covariates**:**

$\Pr_p(Y_i \mid \mathrm{x}_i) = g($Age, Gender, Country of birth (Israel, otherwise).

$\Pr(M_i \mid \mathrm{x}_i; Y_i) = g(Y_i,$ Gender, Country of birth).

$\Pr(y_i \mid \mathrm{x}_i, M_i, Y_i) = g(Y_i,$ Age, Gender).

Model	Covariates	Estimates	S.E. (1,000 Para. BS)
$Pr(y \mid x, Tel, Y)$	Constant	0.286	0.0016
	Y	0.293	0.0017
	Age	0.069	0.0273
	Gender	0.233	0.0013

$Pr(y \mid x, Inter, Y)$	Constant	0.283	0.0016
	Y	0.164	0.0009
	Age	0.102	0.0539
	Gender	0.185	0.0010

Prediction of population prop. of persons with academic degree

$\hat{\bar{Y}}_{(Model)} = \frac{1}{N}\sum_{i=1}^{N}\hat{\Pr}(Y_i \mid \mathrm{x}_i;\hat{\alpha}) \rightarrow$ uses covariates of **all** population units.

$\hat{\bar{Y}}_{(Hajek,Model)} = \sum_{i=1}^{n} w_i\, \hat{\Pr}(Y_i \mid \mathrm{x}_i;\hat{\alpha}) / \sum_{i=1}^{n} w_i \rightarrow w_i = 1/\pi_i$.

$\hat{\bar{Y}}_{(Hajek,True)} = \sum_{i=1}^{n} w_i Y_i^{True} / \sum_{i=1}^{n} w_i \rightarrow$ uses the **true values** from register.

$\hat{\bar{Y}}_{(Hajek,Adj)} = \sum_{i=1}^{n*} \tilde{w}_i y_i / \sum_{i=1}^{n*} \tilde{w}_i \rightarrow \{\tilde{w}_i\} \rightarrow$ weights adjusted for nonresponse.

$\hat{\bar{Y}}_{(Hajek,imp)} = \sum_{i=1}^{n} w_i \tilde{y}_i / \sum_{i=1}^{n} w_i$ **;** $\tilde{y}_i = y_i$ if observed, $\tilde{y}_i =$ **imputed** if not.

Measurement Effect?	$\bar{Y}_{(P)}$ - ***True***	$\hat{\bar{Y}}_{(Model)}$	$\hat{\bar{Y}}_{(Hajek,Model)}$	$\hat{\bar{Y}}_{(Hajek,True)}$	$\hat{\bar{Y}}_{(Hajek,Adj)}$	$\hat{\bar{Y}}_{(Hajek,imp)}$
NO	**0.24**	0.26	0.28	0.25	0.36	0.33
YES		0.25	0.23			

Accurate model-dependent estimates. Testing shows no significant measurement effects. Design-based estimates highly biased.

Concluding remarks- proposed estimation method

1- Does not require the existence of covariates satisfying strong ignorability conditions.

2- Applies to any number of modes.

3- Does not assume that the responses obtained by one of the modes are free of measurement errors.

4- **Nonignorable nonresponse** accounted for.

5- Knowledge of covariates for outside the sample **not required**.

6- **Requires** modelling $f(y_i | M_i, Y_i, \mathrm{x}_i)$, $Pr(M_i | Y_i, \mathrm{x}_i)$ and $f_{\boldsymbol{p}}(Y_i | \mathrm{x}_i)$, but the models **can be tested** using standard test procedures. (**Illustrated in the paper**.)

7- Application to other data sets needed**!!**

Dealing with Proxy Surveys as mode effects

- Proxy surveys are very common in Household (**HH**) surveys- one person of the **HH** responds for all the other members.

- Main motivation is to increase the sample size and possibly reduce nonresponse, because if the designated sampled person cannot be reached, another member of the **HH** is contacted.

- One would expect proxy-responses to be less accurate than self-responses, but this is not always the case. (**Illustrated in the article**).

How to deal with proxy surveys?

We propose to handle proxy surveys by considering them as a **special case** of **mode effects**, with the two main modes defined by **direct-*response*** (person responds about himself) and ***indirect- response*** (information obtained by another person of the **HH**).

Within each of the two primary modes, other modes can be defined, such as the mode of response, known characteristics of the respondent, and **nonresponse**.

❖ Application to **LFS** data in Israel illustrates very good performance. See the article.

Use of administrative records

Supposed to provide timely, accurate data with good coverage, but not always the case.

- Israel’s population register covers all the population residing in Israel, but **15%** of the home addresses are wrong.
- Tax records of businesses obtained with a delay of **2 years**.
- No administrative data on opinions, sentiments, etc.
- If data are timely, accurate and contain all required information, avoids the use of a survey.
- Administrative data often used to strengthen survey estimates by use of **statistical models** or **calibration**.
- Government agencies are often reluctant to transfer data to **NSO’s**, “because” of data protection issues.

Integration (matching) of several administrative records

Desired information often contained in more than one file.

- **Matching problematic** if personal identifiers unknown**;** requires probabilistic algorithms based on information in all the records.
- **Coverage of records** might be different and may not apply to same time periods. **Definitions & accuracy** of information may differ between records.
- Possibly **conflicting information** in different records, **e.g.,** different addresses in different records. (**Major problem** with the use of censuses based on administrative records.)
- Possibly magnified problems of **data protection** after integration.

Use of big data for production of official statistics (OS)

Differences between administrative records and big data:

Both are big!!

- Big data often unstructured, diverse, and appears irregularly (**e.g.,** data obtained from social networks, e-commerce**…**)
- Big data updated dynamically**/**timingly.
- Big data not prepared or maintained for administrative or statistical purposes. It is a **by-product**, not produced for **OS** purposes**!!**
- Big data can cease to appear at any time.
- Big data at risk of data manipulation.

Use of big data for production of official statistics (cont.)

- High dimensionality and extremely large size.
- Possible **coverage/selection bias** (we are talking of **OS**).
- Data accessibility, new legislation**?** Permission by the public**?**
- Possibly increased risks of data disclosure.
- New sampling algorithms (to reduce size and control disclosure)**;**
- Integration of files from multiple sources in different formats appearing at different times.

Shall we really get what we need for official statistics?

Other important issues

Non-representativeness- major concern in use of big data for **OS**.

House sales advertised on the internet do not represent properly all house sales. **Web scraping** for job vacancies does not represent all job vacancies, data from **social media** not representative of **general public**.

No problem when using big data as **predictors** of other variables.

Use **BPP** (Billion Price Project) to predict the **CPI**, **job adverts** to predict **employment** or **job vacancies**, **Satellite images** to predict **crops…**

Requires proper statistical analysis to **identify and test (routinely)** the prediction models.

Accounting for non-representativity of big data

- Big Data may be considered as a special case of a **nonprobability sample (NP)**. Voluntary Internet samples is another example.

Non-repesentativity of nonprobability samples (NP) is a **major concern** in their use for **OS**.

Methods considered in the literature to deal with **NP** samples can be divided into two classes:

1- Integration of **NP** sample with appropriate probability sample (**PS**),

2- Consideration of the **NP** sample on it own. (No data integration.)

Integration of NP samples with PS samples

Several procedures proposed in the literature. Below are two examples:

1- Rivers (2007) proposes to deal with the non-representativeness of a **NP** sample S_{NP} by use of **sample matching**. The approach consists of using a **PS** (reference) sample S_{PS} from the target population ***P***, drawn with known inclusion probabilities $\pi_i = \Pr(i \in S_{PS})$, and then matching to every unit $i \in S_{PS}$ an element ***k*** from the **NP** sample, with similar values of auxiliary (matching) variables $\mathbf{x}$.

Rivers' procedure (cont.)

Denote by $\boldsymbol{x}_i, i=1,...,n$ the $\boldsymbol{x}$ vectors in S_{PS} and by $\tilde{\boldsymbol{x}}_j$ the vectors in S_{NP}.

The unit $k \in S_{NP}$ satisfying $|\tilde{x}_k - x_i| \leq |\tilde{x}_j - x_i| \, \forall j \in S_{NP}$ is chosen as the matched element for unit $i \in S_{PS}$, where |·| is an appropriate distance.

Selecting a matching element for every unit $i \in S_{PS}$ defines a matched sample S_M of size ***n*** of elements from the **NP** sample.

Estimation of population total: $Y = \sum_{i \in P} Y_i \rightarrow \hat{Y}_{SM} = \sum_{k \in S_M} (1/\pi_k)\tilde{y}_k$ **;**

$\tilde{y}_k$= Target y-values measured in S_{NP}, not measured in S_{PS}.

- **Rivers (2007)** established asymptotic properties of estimator.

Integration of NP samples with PS samples (cont.)

2- Kim & Wang (2019) propose the following procedure to account for **non-representativeness** of the **NP** sample:

Assumption: membership of **PS elements** in S_{NP} **known**.

Let $\delta_i = 1(0)$ if $i \in S_{NP}\,(i \notin S_{NP})$. $\boldsymbol{S_{PS}}$ **data:** $\{(\mathrm{x}_i, \delta_i); i = 1,\ldots,n\}$; $\mathbf{x}_i$- model covariates.

Procedure: Model $q_i = \Pr(\delta_i = 1 \,|\, \mathrm{x}_i)$ by use of $S_{PS} \Rightarrow \boldsymbol{\hat{q}_i}$.

Estimate: $\hat{\bar{Y}}_{S_{NP}} = \dfrac{\sum_{i \in S_{NP}} \hat{q}_i^{-1} y_i}{\sum_{i \in S_{NP}} \hat{q}_i^{-1}}$.

- The authors consider also a doubly robust estimator under the assumption of a population regression model.

Remarks on the two proposed procedures described

Neat ideas and apply to any **Y-values, but with important limitations:**

Requires a **PS** with similar sets of $\mathbf{x}$ values in S_{NP} and S_{PS}.

Assume $\Pr(\delta_i = 1 | \mathrm{x}_i, y_i) = \Pr(\delta_i = 1 | \mathrm{x}_i)$; (**noninformative selection**) .

Kim & Wang assume knowledge of membership in S_{NP} of S_{PS} units.

Assume **existence** of $\mathbf{x}$-variables explaining S_{NP}**- membership**.

Rao (2021) reviews many other estimators based on data integration, distinguishing between the case where ***Y*** is observed in both samples and the case where it is only observed in the nonprobability sample.

Accounting for nonrepresentativity without a probability sample

Suppose first a known population model, given by the mean function $E_m(y_i \mid \boldsymbol{x}_i) = h(\boldsymbol{x}_i; \beta); i \in P$ and that the model **holds for** S_{NP}.

Fit the model using the data in S_{NP} to estimate β and predictors $\hat{h}_i(\boldsymbol{x}_i' \hat{\beta})$.

Estimator of Y-total: $\hat{Y}_{S_{NP}} = \sum_{i \in S_{NP}} y_i + \sum_{i \notin S_{NP}} \hat{h}_i(\boldsymbol{x}_i' \hat{\beta})$.

❖ Simple idea but requires that the covariates $\boldsymbol{x}_i$ are known for all populations units and that the population mean model is **known** and **holds** for the nonprobability sample S_{NP}.

See **Rao (2021)** for further discussion of this method.

Accounting for non-representativity (cont.)

Kim and Morikawa (2023) combine a non-ignorable (**informative**) sample selection model with the **empirical likelihood** (**EL**) method.

Let $\delta_i = (1,0)$ be the sampling indicator and denote $\pi_i = \Pr(\delta_i = 1 \mid \boldsymbol{y_i}, \boldsymbol{x}_i)$.

The auxiliary variables $\boldsymbol{x}_i$ are assumed to be **known for all** $i \in U$.

EL Equations: $l(\boldsymbol{p}) = \sum_{i \in S_{NP}} \log(p_i)$,**s.t.** **(1)** $\sum_{i \in S_{NP}} p_i = 1$,

(2) $\sum_{i \in S_{NP}} p_i \pi_i(\boldsymbol{x}_i, y_i) = n / N$ **;** **(3)** $\sum_{i \in S_{NP}} p_i \boldsymbol{x}_i = \bar{\boldsymbol{X}}_U$ (population mean).

↓ ↓

Bias calibration constraint, **Improve efficiency of EL estimator**.

Kim and Morikawa (2023) (cont.)

In practice, the sample selection probabilities $\pi_i = \Pr(\delta_i = 1 | \boldsymbol{y_i}, \boldsymbol{x}_i)$ are unknown. The authors assume therefore a parametric model for the probabilities $\pi_i = \Pr(\delta_i = 1 | \boldsymbol{y_i}, \boldsymbol{x}_i) = g(\boldsymbol{y_i}, \boldsymbol{x}_i; \phi)$, (say, a **logistic model**), and estimate $\hat{\pi}_i = g(\boldsymbol{y_i}, \boldsymbol{x}_i; \hat{\phi})$.

Estimator of population mean with estimated selection prob.

$$\hat{\bar{Y}}_{EL,H-T} = \frac{1}{N} \sum\nolimits_{i \in S_{NP}} \frac{y_i}{\hat{\pi}_i} \textbf{ or } \hat{\bar{Y}}_{EL} = \sum\nolimits_{i \in S_{NP}} \hat{p}_i y_i \,.$$

❖ See the article for details of application of the method and variance estimation. The article contains also discussions by other authors.

Alternative method for inference from nonprobability samples??

A key requirement for the application of the method proposed by **Kim and Morikawa (2023)** is that $\mathbf{x}_i$ **is known** for all $i \in P$.

Here is a possible alternative procedure that does not require this condition. The basic reference is **Pfeffermann & Sverchkov (1999)**.

Population model: $f_p(y_i \mid \mathrm{x}_i) \rightarrow$ model holding for target population outcomes in ***P*** (**"census model"**).

S_{NP} model: $f_{S_{NP}}(y_i \mid \mathrm{x}_i) \rightarrow$ model holding for S_{NP} data.

Denote, as before**;** $\delta_i = 1(0)$ **if** $i \in S_{NP}\,(i \notin S_{NP})$.

Assumption: $\Pr(i \in S_{NP}) > 0\ \forall (i \in P)$.

Alternative method (cont.)

$$f_{S_{NP}}(y_i \mid \mathrm{x}_i) \overset{def}{=} f(y_i \mid \mathrm{x}_i, \delta_i = 1) \overset{\text{Bayes}}{=} \frac{\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(\delta_i = 1 \mid \mathrm{x}_i)}$$

(**) $f_{S_{NP}}(y_i \mid \mathrm{x}_i) = f_p(y_i \mid \mathrm{x}_i)$ **iff** $\Pr(\delta_i = 1 \mid y_i, \mathrm{x}_i) = \Pr(\delta_i = 1 \mid \mathrm{x}_i) \, \forall y_i$.

If **(**) satisfied,** the population and NP sample distributions are the same.

Target *pdf* is $f_p(y \mid \mathrm{x})$**;** observations only available for $f_{S_{NP}}(y \mid \mathrm{x})$.

The two distributions are connected via the **probability link function**, $\Pr(\delta \mid y, \mathrm{x})$.

Alternative method (cont.)

$$f_{S_{NP}}(y_i \mid \mathrm{x}_i) = \frac{\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(\delta_i = 1 \mid \mathrm{x}_i)}$$

Enables estimating **target population pdf** from observations in S_{NP}.

- **No need to know x-values for units not in** S_{NP}.
- $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$ allowed to depend on target variable, **y**. May depend also on other variables ***z***, but **only need** modelling $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$. (May include $\mathbf{z}$ among the $\mathbf{x}$-variables).
- Inference requires modelling $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$ and $f_p(y_i \mid \mathrm{x}_i)$, but **no probability sample required**.
- Use of **Logistic model** for δ_i has some theoretical justification.

Simulation study

Applied same simulation study as in **Kim and Morikawa (2023)**.

1- Generate **5,000** population values as $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$; $x_{1i}, x_{2i} \overset{Inep}{\sim} N(2,1)$; $\varepsilon_i \sim N(0,1)$.

2- $\Pr(\delta_i = 1) = \dfrac{\exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)}$. (Selection probabilities to $\boldsymbol{S_{NP}}$).

3- Repeat Steps 1 and 2, **1,000** times. Overall response rate $\cong$ **50%**.

❖ See true coefficients in next slide.

4- For each simulation estimate the model parameters and the population mean $\bar{Y} = \sum_{i=1}^{5,000} y_i / N$.

Estimation of model coefficients under proposed method

	Population model			Selection probabilities		
	β_0	β_1	β_2	ϕ_0	ϕ_1	ϕ_2
True coefficient	-4	1	1	-2	1	0.5
Mean estimator*	-3.98	1.0	1.0	-1.70	1.22	0.6
Empirical S.E*	.012	.003	.001	0.088	0.037	0.007

* **Mean estimator** and **Empirical S.E.** over **330** simulations.

* “Perfect” estimation of β-coefficients. Less so of ϕ coefficients.

Estimators of finite population means considered:

1- $\hat{\hat{Y}}_{PopXknown} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{x}_i'\hat{\beta}$; **2-** $\hat{\hat{Y}}_{PopXunknown} = \frac{\sum_{i \in S_{NP}} \hat{w}_i y_i}{\sum_{i \in S_{NP}} \hat{w}_i}$; $\hat{w}_i = \frac{1}{\hat{\Pr}(\delta_i = 1)}$.

Estimation of population mean

Method	Bias	Emp. Var X1000	MSE X 1000
Proposed (Pop. x known)*	0.022	0.024	0.508
Proposed (Pop. x unknown)	-0.05	0.88	3.38
Kim & Morikawa	0.01	2.03	2.08
MAR assumption	0.25	1.34	63.84

* Estimation of model coefficients based only on **x**-values in S_{NP}.

* Proposed estimator- **Pop. x known** dominates all other estimators.

* Proposed estimator- **Pop. x unknown** also performs relatively well.
No benchmark constrains**!!** (so far).

* Estimator under **MAR assumption highly biased**.

Model testing

Simulation results are under the correct models.

No direct testing of the population model or the informative selection probabilities is possible, since no data are available from the population distribution and the *y*-values are unknown for units $j \notin S_{NP}$.

However, one can test the distribution $f_{S_{NP}}(y_i \mid \mathrm{x}_i)$, derived from the two models using **classical tests**, since the data in S_{NP} are **known**.

See **Krieger and Pfeffermann (1997)** and **Pfeffermann & Sikov (2011)** for plausible tests.

❖ Rejection of null hypothesis ⇒ at least one of the models is wrong.

Concluding remarks on use of nonprobability samples for OS

- **Non-representativeness** of **NP** samples a major concern.
- Use of **NP** samples for **OS** **not straightforward**.
- Use of **NP** samples for **OS** **inevitable** in the long run. Promises huge advantages, which cannot be ignored.
- The procedures outlined in this presentation to deal with the problem are promising, but only **first** steps.
- Much more theoretical and applied research required**!!**

References mentioned in presentation

De leeuw, E.D. (2018). Mixed-mode: past, present and future. *Survey Research Methods,* **12**, 75–89.

Kim, J.K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, **87**, 177-191.

Kim, J. K. and Morikawa. K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. *Calcutta Statistical Association Bulletin*, **75**. (To appear.)

Krieger, A.M. and Pfeffermann, D. (1997). Testing of Distribution Functions from Complex Sample Surveys. *Journal of Official Statistics,* **13**, 123-142.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *The Journal of Survey Statistics and Methodology* (*JSSAM*), **3**, 425–483.

Pfeffermann, D. and Sverchkov M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhya*, Series *B,* **61**, 166–186.

Pfeffermann, D. and Sikov, A. (2011). Imputation and estimation under non-ignorable nonresponse in household surveys with missing covariate information. Journal of Official Statistics, **27**, 181–209.

Pfeffermann, D. and Preminger, A. (2021). Estimation under Mode Effects and Proxy Surveys, Accounting for Non-ignorable Nonresponse. *Sankhya*, Series A, **83**, 779-813.

Rao, J.N.K. (2021). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya*, Series B, **83**, 242-272.

Rivers, D. (2007). Sampling for web surveys. In *ASA Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, VA, pp. 4127–4134.