Official statistics on population parameters from a nonprobability sample or by integration with a probability sample

Danny Pfeffermann, Daniela Marella and Donato Summa

Third Workshop on Methodologies for official Statistics, Istat, December, 2024

Introduction

- Tightened budgets and constant decrease of response rates in traditional surveys has stimulated research into the use of **non probability** sample data such as administrative records, voluntary internet surveys and other kinds of nonprobability samples.
- A major concern with the use of this kind of data is their nonrepresentativeness, due to possible selection bias, which if not accounted for properly, may bias the inference.

What is official statistics? Why is it important?

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

- **Official Statistics (OS)** is what people hear of almost daily. Unemployment rates, price indexes, poverty measures, house prices, establishment statistics, 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 decisions.

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- **Nonprobability samples (NP); not really implemented routinely yet for OS;** increased pressure on **NSOs** all over the world to **digitise** (**"modernise"**) their data sources.

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

Main issue in the use of nonprobability samples

Non-representativeness- major concern in use of **NP** 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. **Social media** not representative of the general public.

No problem when using the **NP** 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-representativeness of big data

Non-representativeness of nonprobability samples 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 the **NP** sample (S_{NP}) with an appropriate probability sample (S_{PS}),

2- Consideration of the S_{NP} sample on its own. (No data integration.)

Problem considered in this presentation

- In this presentation, we consider the following situation. A large, nonprobability sample S_{NP} with observations on variables ***Y*** and ***X*** is available, but this sample is possibly affected by seletion bias.
- An informative probability sample (S_{PS}), possibly subjected to not missing at random (**NMAR**) nonrespose may also be available, but this sample only contains observations on ***X***.
- The aim is to estimate the joint probability distribution function of the variables ***X*** and ***Y*** and/or the total of ***Y*** in the population.
- We consider alternative methods of integrating the data available in the two samples. We also consider the case where only data from the **nonprobability** sample is available.

Integration of NP samples with PS samples (cont.)

Suppose that the S_{NP} and S_{PS} samples have no units in common.

Following, we apply the **empirical likelihood** approach under which the population distribution is approximated by a **multinomial distribution,** with supports given by the empirical observations in the two samples.

The aim is to estimate $p_i^{XY} = \Pr(X = x_i, Y = y_i)$ of the multinomial distribution and the total $Y = \sum_{i \in U} y_i$.

Denote $p_i^X = \Pr(X = x_i)$ the population probability of ***X***.

Estimation of population probability of X from S_{PS}

Let $I_i^{PS} = 1(0)$ if population unit i is drawn (not drawn) to S_{PS}.

For $i \in S_{PS}$, the sample probability of X is given by,

$$p_{i,PS}^{X} = P(x_i \mid I_i^{PS} = 1) \overset{Bayes}{=} \frac{P(I_i^{PS} = 1 \mid x_i)}{P(I_i^{PS} = 1)} p_i^{X}.$$ (**Informative** sampling.)

- $p_{i,PS}^{X} \neq p_i^{X}$ unless $P(I_i^{PS} = 1 \mid x_i) = P(I_i^{PS} = 1)$.

Next, suppose that the sample S_{PS} is also subject to not missing at random (**NMAR)** nonresponse (see next slide). Let $R_i^{PS} = 1(0)$ if $i \in A$ responds, (does not respond) and denote by R_{PS} the set of the responding units. The response process is assumed to be independent between units.

Estimation of population probability of X from S_{PS} (cont.)

For $i \in R_{PS}$, the probability $P(X = x_i)$ is,

$$p_{i,R_{PS}}^{X} = P(x_i \mid I_i^{PS} = 1, R_i^{PS} = 1) = \frac{P(R_i^{PS} = 1 \mid x_i, I_i^{PS} = 1)}{P(R_i^{PS} = 1 \mid I_i^{PS} = 1)} \; p_{i,PS}^{X}$$

$$= \frac{P(R_i^{PS} = 1 \mid x_i, I_i^{PS} = 1)}{P(R_i^{PS} = 1 \mid I_i^{PS} = 1)} \frac{P(I_i^{PS} = 1 \mid x_i)}{P(I_i^{PS} = 1)} \; p_i^{X} \; (\textbf{NMAR nonresponse})$$

The probability $p_{i,PS}^{X}$ is a function of the corresponding population probability, the conditional probability $P(I_i^{PS} = 1 \mid x_i))$ and the response probabilities $P(R_i^{PS} = 1 \mid x_i, I_i^{PS} = 1)$.

- Unless $P(R_i^{PS} = 1 \mid x_i, I_i^{PS} = 1) = P(R_i^{PS} = 1 \mid I_i^{PS} = 1)$, $p_{i,R_{PS}}^{X} \overset{\text{NMAR}}{\neq} p_{i,PS}^{X} \overset{\text{INFORM}}{\neq} p_i^{X}$.

Estimation of the sampling and response probabilities

The sampling probabilities satisfy $P(I_i^{PS}=1|x_i)=1/E_{PS}(w_{i,PS}|x_i)$; (Pfeffermann and Sverchkov (1999; $w_{i,PS}=1/\pi_{i,PS}$ are the sampling weights in S_{PS}). Thus, assuming that the response is independent of the sample selection, $E_{PS}(w_{i,PS}|x_i)=E_{R_{PS}}(w_{i,PS}|x_i)$ and $P(I_i^{PS}=1|x_i)$ can be estimated by regressing $w_{i,PS}$ against x_i, using the observed data in S_{PS}.

The response probabilities $P(R_i^{PS}=1|x_i,I_i^{PS}=1)$ are unknown and need to be estimated by postulating a parametric model (say, logistic),

$P(R_i^{PS}=1|x_i,I_i^{PS}=1,\rho)=g(x_i;\rho)$ with ρ defining the model parameters.

Estimation of the Probability function of X from S_{PS} (cont.)

Denote $R_{PS,i} = \{x_k \in R_{PS}; x_k = x_i\} \Rightarrow p^X_{k,R_{PS}} = p^X_{i,R_{PS,i}} \; \forall k \in R_{PS,i}; R_{PS} = \cup_i R_{PS,i}$. The *empirical respondents likelihood* for x based on R_{PS} is thus,

$$ERL_{R_{PS}}(p_i^X) = \prod_{i \in R_{PS}} p^X_{i,R_{PS}} = \prod_{i \in R_{PS}} \frac{P(R_i^{PS} = 1 | x_i, I_i^{PS} = 1)}{P(R_i^{PS} = 1 | I_i^{PS} = 1)} \frac{P(I_i^{PS} = 1 | x_i)}{P(I_i^{PS} = 1)} p_i^X ,$$

which depends only on the observed data for the responding units.

The unknown parameters are the population probabilities p_i^X and the response parameters ρ. (The probabilities $P(I_i^{PS} = 1 | x_i)$ are estimated outside the likelihood.)

Estimation based on the nonprobability sample S_{NP}

In S_{NP} we observe **X** and **Y**, but it may be subject to informative selection (depending on both variables). Let I_i^{NP} be the sample indicator. We again apply the **EL** to approximate the population probabilities of (x, y) by a multinomial model with probabilities $p_i^{XY} = \Pr(X = x_i, Y = y_i)$. The sample probabilities in S_{NP} are,

$$p_{i,NP}^{XY} = P(x_i, y_i \mid I_i^{NP} = 1) = \frac{P(I_i^{NP} = 1 \mid x_i, y_i)}{P(I_i^{NP} = 1)} p_i^{XY};$$ (Informative selection.)

$P(I_i^{NP} = 1) = \sum_{i \in NP} P(I_i^{NP} = 1 \mid x_i, y_i) p_i^{XY}$. The probabilities $P(I_i^{NP} = 1 \mid x_i, y_i)$ are modelled parametrically (say, logistic), $P(I_i^{NP} = 1 \mid x_i, y_i; \gamma) = h(y_i, x_i; \gamma)$, with γ defining the model parameters.

Empirical likelihoods based on S_{NP} and $S_{NP} \cup S_{PS}$

The empirical sample likelihood based on S_{NP} is thus,

$$ESL_{NP}(p_i^{XY}) = \prod_{i \in NP} p_{i,NP}^{XY}.$$

We assume that there are no common units in the two samples. (See remark later.) The *empirical likelihood* based on the data in $S_{R_{PS}}$ and S_{NP} is thus the product of the two likelihoods,

$$EL_{R_{PS} \cup S_{NP}} = ERL_{R_{PS}}(p_i^X) ESL_{NP}(p_i^{XY}) = \prod_{i \in R_{PS}} p_{i,R_{PS}}^X \prod_{i \in NP} p_{i,NP}^{XY}.$$

The unknown parameters are the population probabilities p_i^X, p_i^{XY}, the sampling parameters γ and the response parameters ρ.

Maximization of the likelihood

The likelihood is maximized subject to the constraints

$$p_i^X \geq 0,\ p_i^{XY} \geq 0,\ \sum_{i \in R_{PS}} p_i^X = 1,\ \sum_{i \in S_{NP}} p_i^{XY} = 1.$$

The summations are over the different values of $\{x_i\}$ in R_{PS} and the different values of $\{y_i, x_i\}$ in S_{NP}.

The estimation can be further enhanced by adding calibration constraints. In particular, if the population size ***N*** and/or the population mean $\boldsymbol{\mu_x}$ of x are known, add some or all the constraints,

$$\sum_{i \in R_{PS}} x_i p_i^X = \sum_{i \in S_{NP}} x_i p_i^X = \mu_x,\ \sum_{i \in S_{NP}} [P(I_i^{NP} = 1 \mid x_i, y_i; \gamma)]^{-1} = N$$

$$\sum_{i \in R_{PS}} [P(R_i^{PS} = 1 \mid x_i, I_i^{PS} = 1; \rho) \times P(I_i^{PS} = 1 \mid x_i)]^{-1} = N.$$

Estimation of the probabilities P_i^X from both samples

The probabilities $\{p_i^X\}$ can also be estimated from S_{NP}; $\hat{p}_{i,NP}^X = \sum_{\{i;x=x_i\}} \hat{p}_{i,NP}^{XY}$.

This implies two sets of estimates of the probabilities, which can be harmonized as, $\hat{p}_i^X = \lambda \hat{p}_{i,R_{PS}}^X + (1-\lambda)\hat{p}_{i,NP}^X$; $\lambda \in [0,1]$.

A plausible choice is $\lambda = n_A/(n_A+n_B)$. See Marella and Pfeffermann (2023) for other harmonization possibilities.

The final, integrated estimate of p_i^{XY} is, $\hat{p}_i^{XY} = \hat{p}_i^X(\hat{p}_{i,NP}^{XY}/\hat{p}_{i,NP}^X)$.

Estimation of population total *Y*:

$$\hat{Y}_{NP}(1) = N\sum_{i\in S_{NP}} y_i \hat{p}_i^Y \quad \textbf{or} \quad \hat{Y}_{NP}(2) = N\frac{\sum_{i\in S_{NP}} \hat{\Pr}^{-1}(I_i^{NP}=1|\,x_i,y_i)y_i}{\sum_{i\in S_{NP}} \hat{\Pr}^{-1}(I_i^{NP}=1|\,x_i,y_i)}.$$

Further Remarks

1- If some of the units in S_{PS} are also included in S_{NP} with a unique identifier, we can apply the proposed approach to the units in S_{PS}, not included in S_{NP}, after modification of the sampling weights in S_{PS}.

2- We considered integration of the data in S_{PS} and S_{NP}, but we showed how the probabilities $p_i^{XY} = \Pr(X = x_i, Y = y_i)$ can be estimated solely from the nonprobability sample S_{NP}.

3- The proposed approach does not require knowledge of the x-values for all population units but it does assume knowledge of some of their means for the calibration constraints.

Remarks (cont.)

4- The inclusion probabilities $P(I_i^{PS}=1|\,x_i), P(I_i^{NP}=1|\,x_i,y_i)$, and the response probabilities $P(R_i^{PS}=1|\,x_i, I_i^{PS}=1)$ might depend on other variables, but only need to model them as functions of x_i and y_i.

5- The proposed approach is model dependent, but the models can be tested, using classical test statistics. (See later.)

6- The likelihoods considered with the calibration constraints are maximized by application of the profile likelihood, using the function **solNP in R**. See Marella (2023) and Marella and Pfeffermann (2023) for further details.

Other approaches proposed in the literature for sample integration

Rivers (2007) proposes to deal with the **selection bias** to S_{NP} by use of **sample matching**. The approach consists of using a S_{PS} (reference) sample from the target population, drawn with known probabilities $\pi_i = \Pr(i \in S_{PS})$, and then matching to every unit $i \in S_{PS}$ an element ***k*** from S_{NP}, with x_k being the closest to x_i by some distance metric.

Defines a matched probability sample S_M with observations $(x_i, \tilde{y}_i)$ where $\tilde{y}_i$ is the y-value of the matched element x_k, measured in S_{NP}.

Estimation based on S_M using classical survey sampling methods.

- Instead of matching one record, one can match ***k*** nearest records and **select randomly** the matched record out of them. (**kNN)**.

Other approaches for sample integration (cont.)

Kim & Wang (2019) propose the following procedure to correct for the **selection bias** of S_{NP}:

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

Let $\delta_i = 1(0)$ if $i \in S_{NP}$ $(i \notin S_{NP})$. S_{PS} data: $\{(x_i, \delta_i);\ i = 1,...,n\}$.

Procedure: **Model** $q_i = \Pr(\delta_i = 1 \mid x_i; \gamma)$ by use of $S_{PS} \Rightarrow \hat{q}_i = q_i(\hat{\gamma})$.

Estimation of population total based on S_{NP} sample,

$$\hat{Y}_{S_{NP}}(1) = \sum_{i \in S_{NP}} q_i^{-1}(\hat{\gamma}) y_i \quad \textbf{or} \quad \hat{Y}_{S_{NP}}(2) = N \sum_{i \in S_{NP}} q_i^{-1}(\hat{\gamma}) y_i \Big/ \sum_{i \in S_{NP}} q_i^{-1}(\hat{\gamma}).$$

- The last two approaches are suitable for **MAR** selection to S_{NP}. The S_{PS} sample is assumed to be fully observed.

A proposed approach for inference from only S_{NP}

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

Let $\delta_i = (1,0)$ be the sample indicator and denote

$\pi_i(y_i, x_i) = \Pr(\delta_i = 1 | \boldsymbol{y_i}, x_i)$.

- The auxiliary variables 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)** $\sum_{i \in S_{NP}} p_i = 1$**;** $\boldsymbol{p_i}$**=EL probab**.

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

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

Kim and Morikawa approach (cont.)

In practice, the sample selection probabilities $\pi_i(y_i, x_i) = \Pr(\delta_i = 1 | \boldsymbol{y_i}, x_i)$ are unknown. The authors assume a parametric model, $\pi_i = h(\boldsymbol{y_i}, x_i; \gamma)$ and estimate $\hat{\pi}_i = h(\boldsymbol{y_i}, \boldsymbol{x}_i; \hat{\gamma})$ outside the likelihood.

Estimation of population totals:

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

A novel approach for estimating finite population parameters from S_{NP} samples subject to nonignorable selection probabilities, but the assumption that x_i is known for all $i \in U$ is restrictive.

Simulation study

Step 1: Generate a finite population ***U*** of **N=30,000** units from a multinomial distribution where $X=(X_1,X_2,X_3)$ and Y is binary**;** $X_1=(1,2,3,4)$ and (X_2,X_3) are categorical with **4** categories (defined by 4 dummy variables). The number of unknown probabilities p_i^{XY} is thus **128**.

Step 2: Draw independently samples S_{PS} and S_{NP} from *U* by use of Poisson sampling with selection probabilities,

$$\pi_{i,S}=n_S\frac{\exp(\kappa_{1,S}x_{1,i}+\kappa'_{2S}x_{2,i}+\kappa'_{3S}x_{3,i}+\kappa_{y,S}y_i)}{\sum_{j=1}^{N}\exp(\kappa_{1,S}x_{1,j}+\kappa'_{2S}x_{2,j}+\kappa'_{3S}x_{3,j}+\kappa_{y,S}y_j)};\ \boldsymbol{S=S_{PS},S_{NP}}.$$

- $\kappa_{y,S}=0$ when selecting S_{PS}. $\bar{n}(S_{PS})=5000,\ \bar{n}(S_{NP})=20,000$.

Simulation study (cont.)

Step 3: Generate the subset of responding units R_{PS} with response probabilities $P(R_i^{PS}=1|\, x_i, I_i^{PS}=1, \rho)=\text{logit}^{-1}(\rho_0+\rho_1 x_{1,i})$. $\rho_0=0.1, \rho_1=0.2$.

Repeat Steps 2 and 3 **500 times. Response rate≅65-70%**.

We consider **3** models for estimating the selection probability coefficients generating the S_{NP} sample:

Model 1: $\kappa'_{NP}=(\kappa_{1,NP},\kappa'_{2,NP},\kappa'_{3,NP},\kappa_{y,NP})=(0.1,0.05,0.1,0.15,0.2,0.05,0.1,0.15,0.2,0.5)$

Model 2: $\kappa'_{NP}=(\kappa_{1,NP},\kappa'_{2,NP},\kappa'_{3,NP},\kappa_{y,NP})=(0.1,0.05,0.1,0.15,0.2,0,0,0,0,0.5)$

Model 3: $\kappa'_{NP}=(\kappa_{1,NP},\kappa'_{2,NP},\kappa'_{3,NP},\kappa_{y,NP})=(0.1,0,0,0,0,0,0,0,0,0.5)$

Simulation study (cont.)

The corresponding models used for estimation of the selection probability coefficients generating the S_{PS} sample are,

Model 1: $\kappa'_{SP} = (\kappa_{1,SP}, \kappa'_{2,SP}, \kappa'_{3,SP}) = (0.1, 0.05, 0.1, 0.15, 0.2, 0.05, 0.1, 0.15, 0.2)$

Model 2: $\kappa'_{SP} = (\kappa_{1,SP}, \kappa'_{2,SP}, \kappa'_{3,SP}) = (0.1, 0.05, 0.1, 0.15, 0.2, 0, 0, 0, 0)$

Model 3: $\kappa'_{SP} = (\kappa_{1,SP}, \kappa'_{2,SP}, \kappa'_{3,SP}) = (0.1, 0, 0, 0, 0, 0, 0, 0, 0)$

Model 1 is the **correct model** generating the data. Models 2 and 3 are **misspecified working models** used for estimation.

Estimation of population probabilities

i) based only on S_{NP}; **ii)** based on $R_{PS} \cup S_{NP}$; **iii)** Kim & Wang method; **iv)** Rivers' method; **v)** kNN method with $k = \sqrt{n_{NP}}$ (Silverman, 1986).

In method **i)**, the empirical likelihood is maximized under the constraints,

$$\sum_{i \in S_{NP}} [P(I_i^{NP} = 1 | x_{1,i}, x_{2,i}, x_{3,i}, y_i; \gamma)]^{-1} = N, \sum_{i \in S_{NP}} x_{1,i} p_i^{x_1} = \mu_{x_1}, p_i^{XY} \geq 0, \sum_{i \in S_{NP}} p_i^{XY} = 1.$$

In method **ii)** the first constraint is replaced by

$\sum_{i \in R_{PS}} [\Pr(R_i^{PS} = 1 | x_i, I_i^{PS} = 1; \rho) \times \Pr(I_i^{PS} = 1 | x_i)]^{-1} = N$. For method **iii)** we fitted,

$$q_i(\gamma) = P(\delta_i = 1 | x_{i,1}, x_{i,2}, x_{i,3}; \gamma) = \text{logit}^{-1}(\gamma_0 + \gamma_1 x_1 + \gamma_2' \boldsymbol{x}_2 + \gamma_3' \boldsymbol{x}_3);$$

$\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are 4 dummy variables.

Simulation Results

Table 1: Mean and standard deviation (*Sd*) of estimated response coefficients. True values: $\rho_0 = 0.1,\ \rho_1 = 0.2$

Model	$\hat{\rho}_0$	$Sd(\hat{\rho}_0)$	$\hat{\rho}_1$	$Sd(\hat{\rho}_1)$
Model 1	0.09	0.0015	0.163	0.0053
Model 2	0.06	0.0010	0.151	0.0046
Model 3	0.03	0.0012	0.102	0.0033

- The larger the distance between the correct model 1 and the model fitted, the larger the bias of the estimated response model coefficients.

Table 2. Mean Hellinger distance between estimated ($\hat{\boldsymbol{p}}$) and true ($\boldsymbol{p}$) probabilities of (X,Y). $\mathbf{HD(\hat{p},p)} = \sqrt{\frac{1}{2}[\sum_{i=1}^{128}(\sqrt{\hat{p}_i} - \sqrt{p_i})^2]}^{1/2}$.

Models	S_{NP}	$R_{PS} \cup S_{NP}$	Kim & Wang	Rivers	kNN
Model 1	0.069	0.061	0.086	0.089	0.089
Model 2	0.080	0.078	0.087	0.089	0.090
Model 3	0.084	0.085	0.089	0.090	0.089

Small values under the correct model, larger under misspecified models and under **MAR** nonresponse.

Table 3. Means and standard deviations (S_d) of estimated values of κ_{NP} coefficients under Model 1.

κ_{NP}	$\kappa_{1,NP}$	$\kappa_{2,NP}^{1}$	$\kappa_{2,NP}^{2}$	$\kappa_{2,NP}^{3}$	$\kappa_{2,NP}^{4}$	$\kappa_{3,NP}^{1}$	$\kappa_{3,NP}^{2}$	$\kappa_{3,NP}^{3}$	$\kappa_{3,NP}^{4}$	$\kappa_{y,NP}$
True coeff.	0.1	0.05	0.1	0.15	0.2	0.15	0.1	0.15	0.2	0.5
Estimate	0.1	0.04	0.13	0.15	0.19	0.06	0.13	0.12	0.22	0.39
S_d	0.00	0.00	0.010	0.012	0.008	0.001	0.009	0.012	0.02	0.01

Mean estimates generally close to true values.

Model testing

As mentioned before, we can actually test the model fitted. We have data for the responding units and a model fitted to them. We applied the Hosmer and Lemeshow (1980, hereafter H-L) test. To construct the test, the sample is partitioned into ***G*** **≈**equal size groups based on the estimated probabilities of success (*Y*=1). The test statistic is,

$\text{H-L} = \sum_{g=1}^{G} \frac{(o_g - n_g \bar{\mu}_g)^2}{n_g \bar{\mu}_g (1 - \bar{\mu}_g)}$, where o_g is the number of observed successes in group g, n_g is the size of the group and $\bar{\mu}_g$ is the mean of the estimated probabilities of success. $\text{H-L} \sim \chi^2_{G-2}$ (approximately) under the null hypothesis that the model fits the data.

Test results

Table 4: Means of p-values of H-L test statistic, ***G*=10**

Models	S_{NP}	$R_{PS} \cup S_{NP}$	*Kim & Wang*	*Rivers*	kNN
Model 1	0.076	0.079	0.0080	0.008	0.0073
Model 2	0.052	0.059	0.0068	0.0061	0.0072
Model 3	0.022	0.029	0.0056	0.0053	0.0067

Under the correct model, the means are larger than 0.075. For the last three methods, the means are all smaller than 0.01. The test rejects Model 3 with very small p-values. Under Model 2, the means are close to 0.05, but Models 1 and 2 are not very different.

Estimation of population total *Y*:

$$\hat{Y}_{NP}(1) = N \sum_{i \in S_{NP}} y_i \hat{p}_i^{Y} \; ; \; \hat{Y}_{NP}(2) = N \frac{\sum_{i \in NP} \hat{\Pr}^{-1}(I_i^{NP} = 1 | x_i, y_i) y_i}{\sum_{i \in NP} \hat{\Pr}^{-1}(I_i^{NP} = 1 | x_i, y_i)}$$

Table 5. Means of estimates of ***Y*** total and standard deviations based on only S_{NP}. True Total=**13,548**

Models	$\hat{y}_{NP}(1)$	$Sd(\hat{y}_{NP}(1))$	$\hat{y}_{NP}(2)$	$Sd(\hat{y}_{NP}(2))$
Model 1	13750.45	116.84	13680.95	60.86
Model 2	14014.28	111.26	13747.59	64.80
Model 3	14099.847	108.26	13798.75	58.91

Both perform relatively well under correct model, but $\hat{y}_{NP}(2)$ has smaller bias and is much less variable. Also performs OK under 2 and 3.

Table 6. Means of estimates of ***Y*** total and standard deviations based on $R_{PS} \cup S_{NP}$. True Total=**13,548**

Models	$\hat{y}_{NP}(1)$	$Sd(\hat{y}_{NP}(1))$	$\hat{y}_{NP}(2)$	$Sd(\hat{y}_{NP}(2))$
Model 1	13708.04	109.96	13654.46	60.86
Model 2	14004.28	103.80	13757.42	64.92
Model 3	14064.58	98.39	13773.57	58.94

Comparing the results of Tables 5 and 6 shows that integrating the data in S_{NP} and R_{PS} results in lower bias of the estimators under the 3 models, and lower standard deviations of $\hat{y}_{NP}(1)$.

Application to real data from Italy

S_{NP} **sample:** Enterprises belonging to **NACE** (Nomenclature of Economic Activities) having at least **10** employees, selected from the statistical register of **italian** active enterprises (ASIA).

The sample contains information on the following variables: $\mathbf{X}_1$= number of employees; $\mathbf{X}_2$=sales (in classes); $\mathbf{X}_3$=*Nace*; $\mathbf{X}_4$=*geographical area*. Y**=1** if involved in e-commerce; **0** otherwise.

- Reference time period- **2022**. Sample size $n_{S_{NP}} = 51,714$.

Application to real data from Italy (Cont.)

S_{PS} **sample:** "Situation and perspectives of **Italian** enterprises during COVID-19". **Noninformative and not affected by nonresponse??**

Carried out by **Istat** with the aim of assessing the economic situation and specific actions taken by businesses to reduce the economic impacts of the pandemic. The survey was conducted in 3 waves, the last between **16/11-17/12, 2021** (considered in the present study).

The sampling design used to select the sample is two-stage stratified random sampling.

Only enterprises with at least **10** employes and belonging to the **NACE** economic activities are considered (same as S_{NP} sample).

The sample size $n_{S_{PS}} = 19,606$.

Application to real data- initial results

Because of operational limitations, only considered the variable **X**- No. of employees. For further evaluation of the method, we divided the X-values into **k=4** categories:

$$z_1 = [10-49];\ z_2 = [50-99];\ z_3 = [100-249];\ z_4 = 250+.$$

We computed the probability function of ***Z*** by employing the probability function of ***X***; $\hat{p}_k^{Z,S} = \sum_{i \in S_{z,k}} \hat{p}_i^{X,S}$; $S_{z,k} = \{i \in S : x_i \in z_k\}$, **S=** S_{PS} , S_{NP}

Table 7. Estimated proportions in classes Z obtained from S_{PS} and S_{NP} and by harmonization of the two estimates, with $\lambda = n_{PS} / (n_{PS} + n_{NP}) = \mathbf{0.27}$

Z	$n_{S_{NP}}$	$\hat{p}^{Z}_{z,S_{PS}}$	$\hat{p}^{Z}_{z,S_{NP}}$	$\lambda \hat{p}^{Z}_{z,PS} + (1-\lambda)\hat{p}^{Z}_{z,NP}$	ISTAT*
[10-49]	39780	0.876	0.824	0.838	0.863
[50-99]	5692	0.066	0.096	0.088	0.074
[100-250]	3845	0.041	0.057	0.052	0.042
250+	2437	0.018	0.023	0.022	0.021

* The last column contains estimates computed by ISTAT for 2022. The estimates were obtained from a sample of **30,000** enterprises called, *Information and Communication Technology (ICT)*, aiming to measure the degree of digitalization of Italian enterprises.

Note, these estimates are also based on a sample.

Application to real data (cont.)

Table 8. Estimation of selection probabilities to S_{NP} and standard errors*

$$\pi_{i,S_{NP}} = n_{S_{NP}} \frac{\exp(\kappa_{x,S_{NP}} x_i + \kappa_{y,S_{NP}} y_i)}{\sum_{j=1}^{n_{S_{NP}}} \exp(\kappa_{x,S_{NP}} x_j + \kappa_{y,S_{NP}} y_j)}$$

$\kappa_{x,S_{NP}}$	$\kappa_{y,S_{NP}}$
0.15 (0.005)	0.82(0.024)

* Standard errors computed based on parametric bootstrap.

- Coefficients highly significant based on standard t-tests. **Clear indication of biased selection**.

Application to real data (cont.)

For the methods considered in the simulation study, we estimated the probabilities $\hat{\mathrm{P}}(Y=1|\,Z=z)=\sum_{M_{z,1}}\hat{p}_i^{XY}\;/\sum_{M_z}\hat{p}_i^{XY}$;

$M_{z,1}=\{i\in z, y_i=1\},\ M_z=\{i\in z\}.$

Table 9. Estimates of $\Pr(Y=1|\,Z=z)$ and ISTAT probabilities

Z	S_{NP}	$S_{PS}\cup S_{NP}$	*Kim & Wang*	*Rivers*	ISTAT*
[10-49]	0.186	0.181	0.148	0.138	0.211
[50-99]	0.172	0.183	0.128	0.120	0.202
[100-250	0.176	0.179	0.151	0.160	0.210
250+	0.240	0.240	0.213	0.220	0.262

* Same sample as in Table 7.

Other results

1- The **HD** measures of the distance between the estimated probabilities and the ISTAT probabilities shown in Table 9 are **0.044** for the method based on S_{NP}, **0.040** for the method based on $S_{PS} \cup S_{NP}$, **0.100** for Kim & Wang method and **0.108** for Rivers' method.

2- The estimated joint probabilities of (X,Y) obtained from S_{NP} and $S_{PS} \cup S_{NP}$ have been tested by use of the H-L test with **G=10** groups. The p-values are **0.069** and **0.083** respectively.

Concluding remarks on use of nonprobability samples for OS

- **Informative sampling** and **NMAR** nonresponse should always be checked and adjusted for, both with **PS** and **NP** samples.
- **Non-representativeness** of **NP** samples a major concern.
- Use of **NP** samples for **OS not straightforward**.
- Use of **NP** samples for **OS** apparently **inevitable** in the long run. Promises huge advantages, which cannot be ignored.
- The procedures outlined in this presentation to deal with selection bias are promising, but only **first** steps.
- Much more theoretical and applied research required**!!**

References mentioned in presentation

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.)

Marella, D. (2023). Adjusting for selection bias in nonprobability samples by the empirical likelihood approach. *Journal of Official Statistics*, 39, 151-172.

Marella, D. and Pfeffermann, D. (2023). Accounting for Non-ignorable Sampling and Non-response in Statistical Matching. *International Statistical Review*, **91**, 269–293.

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.

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.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.