

# Accounting for non-ignorable sampling and nonresponse In Statistical Matching

**Based on new joint article with Daniela Marella. Appears in:**

*International Statistical Review* (2022); <https://doi.org/10.1111/insr.12524>

## Statement of problem

- + Data for statistical analysis may only be available from different samples, with each sample containing measurements on only some of the variables of interest.
- + **Problem:** generate a fused database containing matched data on all the target variables.
- + In this presentation, I consider the case where the samples are drawn by **informative sampling designs** and are subject to **not missing at random (NMAR)** nonresponse.

## Statement of problem (cont.)

✚ If the sampling and nonresponse processes are ignored, the distribution of the observed data for the responding units can be very different from the distribution of the population data, which may distort the inference process and result in a matched database that **misrepresents** the joint distribution in the population.

✚ Our proposed methodology employs the **empirical likelihood approach** and is shown to perform well in a simulation experiment and when applied to real sample data.

## Informative sampling

Let  $A$  and  $B$  be two independent samples of sizes  $n_A$  and  $n_B$ , selected from a population of  $N$  independent and identically distributed records  $(x_i, y_i, z_i)$ , generated from some joint probability density function (*pdf*)  $f_P(x, y, z; \theta)$ .

The statistical matching problem is that **only**  $(X, Y)$  are observed for the units in  $A$ , and **only**  $(X, Z)$  are observed for the units  $B$ .

We assume that the sampling designs are **informative**, *i.e.*, the sample selection probabilities for  $A$  are correlated with at least some of the variables  $(X, Y)$ , and similarly for the sample  $B$ , implying that even if all the three variables had been observed in the two samples, the joint sample *pdf*  $f_S(x, y, z)$  of the sample data is different from the corresponding population *pdf*,  $f_P(x, y, z)$ .

## Not missing at random nonresponse

Additionally, we assume that the two samples are subject to **not missing at random (NMAR) unit nonresponse**, i.e., the probability to respond likewise depends on the study variables.

The data available to the analyst consist therefore of the sets of **responding units** in  $A$ ,  $(R_A)$  and  $B$ ,  $(R_B)$ . Consequently, even if  $A$  and  $B$  contain information on  $(X, Y, Z)$ , the joint *pdf* of the observed data,  $f_{R_S}(x, y, z)$ , differs from the sample *pdf*  $f_S(x, y, z)$  under complete response, and from the population *pdf*  $f_P(x, y, z)$ ;  $S = A, B$ .

The purpose is to generate a **fused dataset** with joint observations on  $(X, Y, Z)$ , which is representative (similar distribution) of the population distribution from which the samples are taken, and use it for inference.

## The empirical likelihood approach

The empirical likelihood (**EL**) combines the robustness of nonparametric methods with the efficiency of the likelihood approach.

It is essentially the likelihood of an approximation to the true population distribution by a **multinomial distribution**, where the unknown parameters are the **point masses** (“**probabilities**”) assigned to the distinct values. Hence, it does not require specifying a parametric population model, and is thus more robust and often easier to implement.

In what follows we assume that  $X$  is **discrete**, taking  $K$  values with probabilities  $p_k^X = P(X = x_k)$ ;  $\sum_{k=1}^K p_k^X = 1$ , while  $Y, Z$  are **continuous**.

## Conditional independence assumption

A common approach to deal with the statistical matching problem is to assume that  $\mathbf{Y}$  and  $\mathbf{Z}$  are independent given  $\mathbf{X}$ , known as the conditional independence assumption (**CIA**).

Let  $(x_i, y_i, z_i)$  define the values associated with unit  $i$  and denote

$$p_i^X = \Pr(X = x_i), \quad p_i^{Y|X} = \Pr(Y = y_i | X = x_i), \quad p_i^{Z|X} = \Pr(Z = z_i | X = x_i),$$

each having its support in the data observed in the samples.

Under the **CIA**, the joint population multinomial probability of  $(x_i, y_i, z_i)$  is given by

$$p_i^{XYZ} = p_i^X p_i^{Y|X} p_i^{Z|X}.$$

## The sample distributions


Let  $A_k = \{i \in A : x_i = x_k\}$  be the set of sampled units in  $A$  with  $X = x_k$ .

Let  $I_i^A$  be the sample indicator taking the value 1 if unit  $i$  is drawn to the sample  $A$  and 0 otherwise. For  $i \in A_k$  denote,

$$\tau_{i,A}^{XY} = P(I_i^A = 1 | x_i, y_i), \quad \tau_{i,A}^X = P(I_i^A = 1 | x_i) = \sum_{j \in A_k} \tau_{j,A}^{XY} p_j^{Y|X} = \tau_{k,A}^X. \quad \text{Hence,}$$

$$p_{i,A}^{Y|X} = P(y_i | x_i, I_i^A = 1) = \frac{P(I_i^A = 1 | x_i, y_i)}{P(I_i^A = 1 | x_i)} p_i^{Y|X} = \frac{\tau_{i,A}^{XY} p_i^{Y|X}}{\sum_{j \in A_k} \tau_{j,A}^{XY} p_j^{Y|X}}. \quad (1)$$

$$p_{k,A}^X = P(x_k | I_i^A = 1) = \frac{P(I_i^A = 1 | x_k)}{P(I_i^A = 1)} p_k^X = \frac{\tau_{k,A}^X p_k^X}{\sum_{j=1}^K \tau_{j,A}^X p_j^X}. \quad (2)$$

 Under **informative sampling**, the sample probabilities **(1)** and **(2)** are **different** from the corresponding **population probabilities**  $p_i^{Y|X}, p_k^X$ .



## Independence under the sample distribution

Even though the sample models are different from the corresponding population models, it is shown in **Pfeffermann *et al.* (1998)** that if the population values are independent under the population model, under mild conditions they are asymptotically independent under the sample model, when the population size increases but the sample size remains fixed.

This permits approximating the sample likelihood by the product of the sample likelihoods over the corresponding sample observations.

Hence, for sufficiently large populations, the sample EL (**ESL**), based on the observed data in  $A$  is,

$$ESL_{Obs}^A = \prod_{k=1}^K (p_{k,A}^X)^{n_{k,A}^X} \prod_{i \in A_k} p_{i,A}^{Y|X} ; n_{k,A}^X = \#(A_k) \quad (3)$$

An analogous expression holds for the **ESL** based on the data in  $B$ .

## The log(ESL) based on the sample $A \cup B$ (under the CIA)

$$\begin{aligned}
 \log(ESL_{Obs}^{A \cup B}) &= \sum_{i \in A_k} \log(\tau_{i,A}^{XY} p_i^{Y|X}) - n_{k,A}^X \log \left( \sum_{i \in A_k} \tau_{i,A}^{XY} p_i^{Y|X} \right) + \sum_{k=1}^K n_{k,A}^X \log(\tau_{k,A}^X p_k^X) \\
 &\quad - \sum_{k=1}^K n_{k,A}^X \log \left( \sum_{j=1}^K \tau_{j,A}^X p_j^X \right) + \sum_{i \in B_k} \log(\tau_{i,B}^{XZ} p_i^{Z|X}) - n_{k,B}^X \log \left( \sum_{i \in B_k} \tau_{i,B}^{XZ} p_i^{Z|X} \right) + \quad (4) \\
 &\quad + \sum_{k=1}^K n_{k,B}^X \log(\tau_{k,B}^X p_k^X) - \sum_{k=1}^K n_{k,B}^X \log \left( \sum_{j=1}^K \tau_{j,B}^X p_j^X \right).
 \end{aligned}$$

- ✚ The unknown parameters in (4) are the probabilities  $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ .
- ✚ The sampling probabilities in  $A$  and  $B$  may depend on many unobserved variables but by definition of the sample *pdf*, one only needs to model the probabilities  $P(I_i^A = 1 | x_i, y_i)$  and  $P(I_i^B = 1 | x_i, z_i)$ .

## Estimation of the conditional sampling probabilities

The probabilities  $\tau_{i,A}^{XY} = P(I_i^A = 1 | x_i, y_i) = 1 / E_A(w_{i,A} | x_i, y_i)$  (**Bayes**) and  $\tau_{i,B}^{XZ} = 1 / E_B(w_{i,B} | x_i, z_i)$  can be estimated outside the likelihood by regressing the sample weights  $w_{i,A} = 1 / \pi_{i,A}$ , ( $w_{i,B} = 1 / \pi_{i,B}$ ) against  $(x_i, y_i)$ ,  $[(x_i, z_i)]$ , using the observed data in  $A$  and  $B$ .

The **ESL** estimators of the unknown probabilities are obtained by maximizing the loglikelihood **(4)**, subject to the constraints,

$$p_k^X \geq 0, p_i^{Y|X} \geq 0, p_i^{Z|X} \geq 0, \sum_{k=1}^K p_k^X = 1, \sum_{j \in A_k} p_j^{Y|X} = 1, \sum_{j \in B_k} p_j^{Z|X} = 1.$$

## Estimators of the unknown probabilities

The estimators of the unknown probabilities are:

$$\hat{p}_{k,A}^X = [n_{k,A}^X (\tau_{k,A}^X)^{-1}] / \sum_{j=1}^K [n_{j,A}^X (\tau_{j,A}^X)^{-1}], \quad \hat{p}_{k,B}^X = [n_{k,B}^X (\tau_{k,B}^X)^{-1}] / \sum_{j=1}^K [n_{j,B}^X (\tau_{j,B}^X)^{-1}] \quad (5)$$

$$\hat{p}_i^{Y|X} = (\tau_{i,A}^{XY})^{-1} / \sum_{j \in A_k} (\tau_{j,A}^{XY})^{-1}, \quad \hat{p}_i^{Z|X} = (\tau_{i,B}^{XZ})^{-1} / \sum_{j \in B_k} (\tau_{j,B}^{XZ})^{-1},$$

where  $\hat{p}_{k,A}^X, \hat{p}_{k,B}^X$  are the estimates of  $p_k^X$  obtained from the two samples.

The estimates  $\hat{p}_{k,A}^X, \hat{p}_{k,B}^X$  can be harmonized into a unique estimate  $p_k^X$  by a linear combination of the two estimates,

$$\hat{p}_k^X = \lambda \hat{p}_{k,A}^X + (1 - \lambda) \hat{p}_{k,B}^X; \quad \lambda \in [0, 1].$$

A plausible choice is  $\lambda = n_A / (n_A + n_B)$ . See the article for other choices.

## Adding calibration constraints when maximizing the ESL

When population means of variables measured in the sample A and/or in B are known, they can be added to the constraints of the **ESL**. The following calibration constraints may be added, depending on data

availability:  $\sum_{k=1}^K p_k^X x_k = \mu_X$ ,  $\sum_{k=1}^K p_k^X \sum_{i \in A_k} p_i^{Y|X} y_i = \mu_Y$ ,  $\sum_{k=1}^K p_k^X \sum_{i \in B_k} p_i^{Z|X} z_i = \mu_Z$ ,

where  $\mu_X, \mu_Y, \mu_Z$  are the known population means of  $X, Y, Z$ , respectively.

✚ In the empirical study we use the constraint  $\sum_{k=1}^K p_k^X x_k = \mu_X$ .

## Generation of a fused data set

Once the probabilities  $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$  governing the population multinomial model have been estimated, a fused data set with joint observations  $(x, y, z)$  is constructed as follows:

- (i)** Generate  $\tilde{n}$  observations taking values  $(x_1, x_2, \dots, x_K)$  with probabilities  $(\hat{p}_1^X, \hat{p}_2^X, \dots, \hat{p}_K^X)$ ;
- (ii)** For  $i = 1, \dots, \tilde{n}$ ,  $k = 1, \dots, K$ , draw a value  $\tilde{y}_i$  from the estimated probability function  $\hat{p}_i^{Y|X}$ , taking the values  $(y_1^k, y_2^k, \dots, y_{n_{k,A}^X}^k)$  with probabilities  $(\hat{p}_1^{Y|x_k}, \hat{p}_2^{Y|x_k}, \dots, \hat{p}_{n_{k,A}^X}^{Y|x_k})$ , where  $n_{k,A}^X = \#\{i \in A : x_i = x_k\}$ .
- (iii)** Apply a similar procedure for drawing values  $\tilde{z}_i$  from the estimated probability function  $\hat{p}_i^{Z|X}$ .

## Comments

- ✚ The consistency of the estimators of the model parameters guarantees that for sufficiently large sample sizes  $n_A$ ,  $n_B$ , the fused data set can be considered as being generated from the joint population *pdf*.
- ✚ Even under the **CIA**, It is not correct to only impute the missing values in the two samples because under informative sampling, the observed  $(x, y)$  values in  $A$  are not representative of the population  $(x, y)$  values. The same holds for the sample  $B$ .

## The EL under informative sampling and NMAR nonresponse

So far we basically assumed full response. In what follows we assume that additionally to informative sampling, the samples  $A$  and  $B$  are subject to **NMAR** nonresponse, by which the response probabilities depend in some stochastic way on the outcome variables of interest.

Let  $R_i^A$  define the response indicator and  $R_A$  denote the set of responding units in  $A$ , of size  $r_A$ . The response process is assumed to be independent between units.



## The EL under informative sampling and NMAR nonresponse (cont.)

Let  $\rho_{i,A}^X = P(R_i^A = 1 | x_i, I_i^A = 1)$ . By **Bayes rule**, for  $i \in A_k$

$$p_{k,R_A}^X = P(x_k | I_i^A = 1, R_i^A = 1) = \frac{P(R_i^A = 1 | x_k, I_i^A = 1)}{P(R_i^A = 1 | I_i^A = 1)} p_{k,A}^X = \frac{\tau_{k,A}^X \rho_{k,A}^X p_k^X}{\sum_{j=1}^K \tau_{j,A}^X \rho_{j,A}^X p_j^X}$$

$$p_{i,R_A}^{Y|X} = P(y_i | x_k, I_i^A = 1, R_i^A = 1) = \frac{P(R_i^A = 1 | x_k, y_i, I_i^A = 1)}{P(R_i^A = 1 | x_k, I_i^A = 1)} p_{i,A}^{Y|X} = \frac{\tau_{i,A}^{XY} \rho_{i,A}^{XY} p_i^{Y|X}}{\sum_{i \in R_{A,k}} \tau_{i,A}^{XY} \rho_{i,A}^{XY} p_i^{Y|X}}$$

$R_{A,k} = \{i \in R_A : x_i = x_k\}$  defines the group of respondents in  $A$  with  $X = x_k$ ,

$$\rho_{k,A}^X = P(R_i^A = 1 | x_k, I_i^A = 1) = E_A(R_i^A | x_k, I_i^A = 1) = \sum_{i \in R_{A,k}} \rho_{i,A}^{XY} p_{i,A}^{Y|X},$$

$$\rho_{i,A}^{XY} = P(R_i^A = 1 | x_k, y_i, I_i^A = 1) = E_A(R_i^A | x_k, y_i, I_i^A = 1).$$

## The EL under informative sampling and NMAR nonresponse (cont.)

The respondents models are functions of the corresponding population model, the conditional expectations of the sampling weights,  $\tau_{i,A}^{XY} = P(I_i^A = 1 | x_i, y_i) = 1 / E_A(w_{i,A} | x_i, y_i)$ , and the response probabilities  $\rho_{i,A}^{XY} = P(R_i^A = 1 | x_k, y_i, I_i^A = 1)$ . Assuming that the response is independent of the sample selection,  $E_{R_A}(w_{i,A} | x_i, y_i) = E_A(w_{i,A} | x_i, y_i)$ , in which case the probabilities  $P(I_i^A = 1 | x_i, y_i)$  can be estimated by regressing  $w_{i,A}$  against  $(x_i, y_i)$ , using the observed data in  $A$ , and similarly for the sample  $B$ , same as for the case of full response.

## The empirical respondents' likelihood

With straightforward modification of the notation, similar expressions are obtained for the model holding for the responding units in **B**. Thus, the *empirical respondents' likelihood* (**ERL**) for the sample  $A \cup B$  is given by,

$$ERL_{Obs}^{A \cup B} = \prod_{k=1}^K (p_{k,R_A}^X)^{r_{k,A}^X} \prod_{i \in R_{A,k}} p_{i,R_A}^{Y/X} \prod_{k=1}^K (p_{k,R_B}^X)^{r_{k,B}^X} \prod_{i \in R_{B,k}} p_{i,R_B}^{Z/X} . \quad (6)$$

- ✚ The likelihood only uses the observed data for the responding units in the two samples.

## Modelling the response probabilities

The response probabilities are unknown and need to be estimated from the available data. Since no "response weights" are known, parametric models for the response probabilities need to be postulated. For example,

$$P(R_i^A = 1 | x_i, y_i, I_i^A = 1) = g_A(\gamma_{0,A} + \gamma_{x,A}x_i + \gamma_{y,A}y_i),$$

$$P(R_i^B = 1 | x_i, z_i, I_i^B = 1) = g_B(\gamma_{0,B} + \gamma_{x,B}x_i + \gamma_{z,B}z_i),$$

for some functions  $g_A, g_B$ , with unknown parameters  $\gamma_A = (\gamma_{0,A}, \gamma_{x,A}, \gamma_{y,A})$ ,  $\gamma_B = (\gamma_{0,B}, \gamma_{x,B}, \gamma_{z,B})$ .

✚ Modelling the response probabilities by the **logit** or **probit** functions is common, but in our case the probabilities depend also on the outcome variables, which is different from the familiar "propensity scores" approach, under which the response probabilities only depend on the observed covariates, which is in common use under **MAR** nonresponse.

✚ The unknown vector parameters,  $\gamma_A, \gamma_B$ , indexing the response models in the two samples are estimated as part of the maximization of the likelihood. Thus, one needs to maximize the likelihood with respect to a larger set of parameters;  $[\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}, \gamma_A, \gamma_B]$ , for all  $k$  and  $i$ .

### Estimation of all the unknown parameters

Suppose that the probabilities  $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$  have been estimated. The (profile) likelihood” of  $\gamma_A, \gamma_B$  is  $G(\gamma_A, \gamma_B) = ERL_{Obs}^{A \cup B}(\gamma_A, \gamma_B | \hat{p}_k^X, \hat{p}_i^{Y|X}, \hat{p}_i^{Z|X})$ , and it is maximized with respect to  $(\gamma_A, \gamma_B)$ , yielding the estimators,

$$(\hat{\gamma}_A, \hat{\gamma}_B) = \arg \max_{(\gamma_A, \gamma_B)} ERL_{Obs}^{A \cup B}(\gamma_A, \gamma_B | \hat{p}_k^X, \hat{p}_i^{Y|X}, \hat{p}_i^{Z|X}).$$

## Estimation of all the unknown parameters (cont.)

Substituting the estimates into the likelihood **(6)** and maximizing with respect to the unknown sets of probabilities, yields,

$$(\hat{p}_k^X, \hat{p}_i^{Y|X}, \hat{p}_i^{Z|X}) = \arg \max_{(p_k^X, p_i^{Y|X}, p_i^{Z|X})} ERL_{Obs}^{A \cup B}(p_k^X, p_i^{Y|X}, p_i^{Z|X}; \hat{\gamma}_A, \hat{\gamma}_B). \quad (7)$$

The procedure is continued iteratively until convergence.

✚ The models for the response probabilities can be tested by testing the estimated models  $\hat{p}_{i,R_A}^{Y|X}$  and  $\hat{p}_{i,R_B}^{Z|X}$  for the observed data, using standard goodness of fit tests.

✚ Once the probabilities of the population multinomial models  $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$  are estimated, a fused data set with observations  $(x, y, z)$  is constructed, following a similar procedure to what was described before.

## Uncertainty in statistical matching when the CIA does not hold

✚ So far, we assumed that the population *pdf* satisfies the **CIA** but clearly, this need not be the case. Denote by  $F_p(y, z | x_k)$  the joint cumulative population distribution function (**cdf**) of  $(Y, Z)$  given  $X = x_k$ , and by  $F_p(y | x_k)$ ,  $G_p(z | x_k)$  the corresponding marginal *cdfs*.

Unless under additional assumptions, the only valid statement regarding  $F_p(y, z | x_k)$  is that it lies in the set  $\Omega_p^k$  of all joint distributions having **marginal cdfs**  $F_p(y | x_k)$ ,  $G_p(z | x_k)$ . **For known**  $F_p(y | x_k)$ ,  $G_p(z | x_k)$ ,

$$L[F_p(y | x_k), G_p(z | x_k)] \leq F_p(y, z | x_k) \leq U[F_p(y | x_k), G_p(z | x_k)],$$

$$U[F_p(y | x_k), G_p(z | x_k)] = \min[F_p(y | x_k), G_p(z | x_k)],$$

$$L[F_p(y | x_k), G_p(z | x_k)] = \max[0, F_p(y | x_k) + G_p(z | x_k) - 1].$$

✚ These are known as the **Fréchet bounds**.

## Uncertainty in statistical matching (cont.)

A natural pointwise uncertainty measure is the length of the interval  $\{L[\dots], U[\dots]\}$ . For  $X = x_k$ , the measure is,

$$\Delta_p^k = \int_{\mathfrak{R}^2} \{U[F_p(y | x_k), G_p(z | x_k)] - L[F_p(y | x_k), G_p(z | x_k)]\} dF_p(y | x_k) dG_p(z | x_k)$$

An overall, averaged measure is,  $\Delta_p = \sum_{k=1}^K \Delta_p^k P_k^X$ .

Denote,  $\Upsilon_{k,R_A} = (y_1^k, y_2^k, \dots, y_{r_{k,A}^X}^k)$ ,  $\Gamma_{k,R_B} = (z_1^k, z_2^k, \dots, z_{r_{k,B}^X}^k)$ .

The pointwise measure is estimated as,

$$\hat{\Delta}_p^k = \frac{1}{r_{k,A}^X r_{k,B}^X} \sum_{y \in \Upsilon_{k,R_A}} \sum_{z \in \Gamma_{k,R_B}} [U(\hat{F}_p(y | x_k), \hat{G}_p(z | x_k)) - L(\hat{F}_p(y | x_k), \hat{G}_p(z | x_k))]$$

The overall uncertainty measure is estimated as,  $\hat{\Delta}_p = \sum_{k=1}^K \hat{\Delta}_p^k \hat{P}_k^X$ .



## Narrowing the bounds by use of external information

The Fréchet bounds are narrowed when additional information is available.

Suppose that it is known that conditionally on  $X = x_k$ , some function of  $(Y, Z)$  satisfies  $a_k \leq c_k(y, z) \leq b_k$ . The class of plausible *pdfs* is now,

$$\Omega_{p,c}^k = \{F_p(y, z | x_k) : F_p(y, \infty | x_k) = F_p(y | x_k), F_p(\infty, z | x_k) = G_p(z | x_k), \\ a_k \leq c_k(y, z) \leq b_k\}$$

✚ In our empirical study we used the constraint  $Y \leq Z$ . With this constraint, the Fréchet bounds (4.1)-(4.2) are,

$$U_c[F_p(y | x_k), G_p(z | x_k)] = \min[F_p(y | x_k), F_p(z | x_k), G_p(z | x_k)]$$

$$L_c[F_p(y | x_k), G_p(z | x_k)] = \max[0, F_p(y | x_k) + G_p(z | x_k) - 1, \min(F_p(y | x_k), F_p(z | x_k)) + G_p(z | x_k) - 1]$$

By choosing a *matching distribution* from this class, the estimated uncertainty measure  $\hat{\Delta}_{p,c}$  provides an upper bound for the matching error.

The statistical matching problem consists now of choosing a *matching distribution* from the class.

### Choosing a matching distribution

Conti *et al.* (2016) proposed a procedure for choosing a *pdf* in the class **(11)**, based on **Iterative Proportional Fitting (IPF)**. The procedure consists of the following steps:

## Choosing a matching distribution (cont.)

**Step 1:** Discretize  $Y$  and  $Z$  by grouping their ascending values in pre-defined classes. Conditional on  $X = x_k$ , the range of  $Y$  is divided into  $h_k$  adjacent intervals  $I_1^{Y|x_k}, \dots, I_h^{Y|x_k}, \dots, I_{h_k}^{Y|x_k}$ , where  $I_h^{Y|x_k} = [y_{h-1}, y_h]$ , with  $y_0 = \min y_i$ ,  $y_h = \max y_i$ . Similar notation applies to the variable  $Z$ ;  $I_g^{Z|x_k} = [z_{g-1}, z_g]$  for  $g = 1, \dots, g_k$ . For  $X = x_k$ , denote by  $Y_{d,h}$  ( $Z_{d,g}$ ) the **midpoints** in each interval. Let  $\{C^k\}$  be the contingency table defined by the  $h_k g_k$  values  $\Upsilon^{YZ|x_k} = [(y_{d,1}, z_{d,1}), \dots, (y_{d,h}, z_{d,g}), \dots, (y_{d,h_k}, z_{d,g_k})]$ , with cell probabilities  $(p_{11}^{Y_{d,k}Z_{d,k}|x_k}, \dots, p_{hg}^{Y_{d,k}Z_{d,k}|x_k}, \dots, p_{h_k g_k}^{Y_{d,k}Z_{d,k}|x_k})$ .

**A separate contingency table is defined for each  $x_k$ .**

The constraint  $a_k \leq c_k(y, z) \leq b_k$  on the support of  $(Y, Z) | x_k$  is applied to the values  $(Y_{d,h}, Z_{d,g})$ , resulting in cells with **structural zeroes**.

## Choosing a matching distribution (cont.)

**Step 2:** For  $X = x_k$ , the marginal probabilities  $p_{h.}^{Y_{d,h}|x_k}$ ,  $p_{.g}^{Z_{d,g}|x_k}$  in  $\{C^k\}$ , **i.e.**, the probabilities that  $Y_{d,h}$  and  $Z_{d,g}$  take the values  $y_{d,h}$ ,  $z_{d,g}$ , are

$$\text{estimated as, } \hat{p}_{h.}^{Y_{d,k}|x_k} = \sum_{i=1}^{r_{k,A}^X} \hat{p}_i^{Y|x_k} I(y_i^k \in I_h^{Y|x_k}), \quad \hat{p}_{.g}^{Z_{d,k}|x_k} = \sum_{i=1}^{r_{k,B}^X} \hat{p}_i^{Z|x_k} I(z_i^k \in I_g^{Z|x_k}),$$

where  $\hat{p}_i^{Y|x_k}$ ,  $\hat{p}_i^{Z|x_k}$  are the MLE of the **ERL**.

**Step 3:** Once the contingency table has been defined, the midpoints  $(Y_{d,h}, Z_{d,g})$  are checked to identify cells in  $\{C^k\}$ , which do not satisfy the constraint  $a_k \leq c_k(y_{d,h}, z_{d,g}) \leq b_k$ . These cells define **structural zeroes**.

The **IPF** initial cell probabilities are,  $p_{hg}^{0, Y_{d,k} Z_{d,k} | x_k} = \delta_{hg} \hat{p}_{h.}^{Y_{d,k}|x_k} \hat{p}_{.g}^{Z_{d,k}|x_k}$  where  $\delta_{hg} = 1$  for cells not containing structural zeroes and  $\delta_{hg} = 0$  otherwise.

## Constructing a fused data set

A fused data set for  $(X, Y, Z)$  is constructed from the estimated matching distribution as follows:

**(I)** Generate  $\tilde{n}$  observations  $\tilde{x}_i$  from the estimated distribution of  $X$  taking values  $(x_1, x_2, \dots, x_K)$  with probabilities  $(\hat{p}_1^X, \hat{p}_2^X, \dots, \hat{p}_K^X)$ . Denote by  $\tilde{n}_k^X$  the number of observations with  $\tilde{x}_i = x_k$ ;

**(II)** For each observation  $x_i$ ,  $i = 1, \dots, \tilde{n}_k^X$ , draw independently  $\tilde{n}_k^X$  pairs  $[(y_{d,1}, z_{d,1}), \dots, (y_{d,h}, z_{d,g}), \dots, (y_{d,h_k}, z_{d,g_k})]$ , with cell probabilities  $(\hat{p}_{11}^{Y_{d,k}Z_{d,k}|x_k}, \dots, \hat{p}_{hg}^{Y_{d,k}Z_{d,k}|x_k}, \dots, \hat{p}_{h_k g_k}^{Y_{d,k}Z_{d,k}|x_k})$ , computed by the **IPF**.

## Application: matching of household income and expenditure

### Samples and sampling designs

We applied the proposed procedure to two real survey data **in Italy**.

A survey collecting information on households' income and wealth (**SHIW**) is conducted by **Banca d'Italia**. Information on consumption expenses is provided by the Household Budget Survey (**HBS**), conducted by **ISTAT**.

✚ This constitutes a serious problem since household data on both income and expenditure are required by policy makers for analyzing the impact of policy strategies. Statistical matching attempts to combine the data obtained from the two different, non-overlapping surveys, drawn from the same target population.

## Samples and sampling designs (cont.)

**SHIW** is drawn in two stages, with municipalities as the primary sampling units and households (**HH**) as the secondary sampling units. We used the **2010** survey, which consists of **387** municipalities drawn with probabilities proportional to size and **7,951 HHs** sampled by simple random sampling. The **HH** income is defined as the combined disposable annual income of all the people living in the **HH** (hereafter **Y**).

The **HBS** uses a similar sampling design and collects detailed information on socio-demographic characteristics and expenditures (hereafter **Z**) on a disaggregated set of commodities. Here again, we use the **2010** survey, which consist of **470** municipalities and **22,227 HHs**.

## Accounting for nonresponse

**SHIW** and **HBS** suffer from low response rates, about **62%** in both samples. It is quite clear that the nonresponse is explained, at least in part, by the size of the HH and the income, or expenditure. The larger the HH, the more possibilities exist to find a contact person for an interview. In addition, HH consisting of only one or two elder people, often tend not to participate in surveys. Furthermore, as often reported in the literature, the response probability tends to decrease as the HH income or expenditure increase. In order to obtain a response rate of about **62%**, we computed the response probabilities in the two samples by use of the logistic models defined before, with coefficients  $(\gamma_{x,A}, \gamma_{y,A}) = (0.2, -0.002)$ ,  $(\gamma_{x,B}, \gamma_{z,B}) = (0.2, -0.003)$ .



## Choosing a matching variable

As stated before, statistical matching is usually based on a set of variables measured in all the data sources (the **X** variables). We considered three variables as plausible candidate matching variables: household size (**hsize**=1,2,3,4+), area of residence, and occupational status.

After a thorough analysis, we found that the **hsize** is the best matching variable, with uncertainty measure  $\hat{\Delta}_{p,c} = \mathbf{0.11}$ , and it remains approximately the same when including all the three matching variables in the analysis; ( $\hat{\Delta}_{p,c} = \mathbf{0.107}$ ). For applying the methodology, we added the

calibration constraint  $\sum_{k=1}^K p_k^X x_k = 2.4$ , (hereafter **C-C**), where **2.4** is the average size of households in **2010**, as published in the **ISTAT** website.

## Results when matching the two surveys

The following table displays 4 different estimates of the probabilities  $\{p_k^X\}$  for the 4 size values (**hsize**=1,2,3,4+).

<b>hsize</b>	$p_k^X$	$\hat{p}_{k,1}^X$	$\hat{p}_{k,1C}^X$	$\hat{p}_{k,2C}^X$	$\hat{p}_{k,2CM}^X$	$n_{k,A}^X$	$n_{k,B}^X$	$r_{k,A}^X$	$r_{k,B}^X$
<b>1</b>	0.284	0.260	0.264	0.276	0.276	5851	1989	3194	1074
<b>2</b>	0.276	0.293	0.293	0.281	0.280	6292	2522	3783	1504
<b>3</b>	0.209	0.210	0.208	0.200	0.205	4758	1589	3069	1028
<b>4+</b>	0.232	0.238	0.233	0.243	0.239	5326	1851	3730	1258

Computing the Hellinger distance  $HD = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^K \left( \sqrt{\hat{p}_k^X} - \sqrt{p_k^X} \right)^2}$ , we find that for

$\hat{p}_{k,1}^X$ , **HD= 0.023**. For  $\hat{p}_{k,1C}^X$ , **HD= 0.018**. For  $\hat{p}_{k,2C}^X$ , **HD= 0.012** and for  $\hat{p}_{k,2CM}^X$  **HD= 0.009**. So, the proposed procedure yields overall the **best estimates**.

## Results when matching the two surveys (cont.)

We also estimated the *pdfs*  $\{p_i^{Y|X}, p_i^{Z|X}\}$  **under the CIA**, both when ignoring the sampling designs and nonresponse and when accounting for them, imposing the **C-C** in both cases. Next, we generated a fused data set of size  $\tilde{n} = 10,000$ . The (weighted) correlations  $cor_{XY}, cor_{XZ}$  in the original samples are **0.38** and **0.31**. In the fused data set, the correlations are **{0.34, 0.28}** when ignoring the sampling designs and nonresponse, and **{0.38, 0.32}** when accounting for them. The correlation between the generated values of  $Y$  and  $Z$  when ignoring the sampling designs and nonresponse in the estimation of the probabilities  $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$  is **0.08**. The correlation increases to **0.13** when both processes are accounted for. When assuming the **CIA**, the correlation computed from the original samples is  $cor_{YZ}^{CIA} = cor_{XY} cor_{XZ} = \mathbf{0.12}$ .

**SHIW** contains also some recall questions for constructing an approximate measure of total expenditure. The correlation in the SHIW sample between income and expenditure is **0.65**. Thus, the fused data set constructed under the **CIA** misrepresents the joint population distribution of  $(Y, Z)$ .

Consequently, we estimated instead a *matching distribution* for income and expenditure by assuming the class of plausible distributions, with the added constraints  $Y \geq Z$  and the **C-C**, applying the **IPF**. Next, we used the estimated joint distribution for generating  $\tilde{n} = 10,000$  values  $(x_i, y_i, z_i)$ .

The correlation between the imputed values of  $Y$  and  $Z$  under the **CIA** is **0.12**. The correlation increases to **0.55** by use of the **IPF**. The correlation in the **SHIW** sample is **0.65**, but the expenditure is not well observed in **SHIW**. (Suffers from significant under-reporting of about **30%**).

**Intermediate conclusion:** The proposed procedure seems to work well!!!

**THANK YOU**