CALIBRATION, BALANCED SAMPLING WITH APPLICATIONS TO NON-RESPONSE

Jean-Claude DEVILLE
ENSAI/Crest
Laboratoire de Statistique d’Enquête
GENERALITIES

1- Universe, label, sample and sample design, inclusion probabilities

2- Unbiased estimation for a total

3- Practical Sampling

→ Decomposition of a sampling plan: Stratification and multi-stage designs

→ Units sampling plans

4- Non-linear Estimation: Plug-in Principle (or substitution)

→ Examples of non-linear indicators to be estimated

→ Plug-in principle for point estimation

→ Linearization for variance approximation with examples

→ Linearization variance estimation

5- Representativeness à la Hajek
BALANCED SAMPLING

1-Entropy and Poisson Sampling
2-Variance, Variance Approximation and Variance Estimation and Extensions

2-1 Necessity of an approximation
2-2 Properties of a good approximation

3-Exact or approximate balance?

4-The Cube Method
4-1 The Cube set-up
4-2 Flight (iterative) phase
4-3 Landing (round-up) phase

5-Variance and Variance Estimation
5-1 How to get maximum entropy?
5-2 Ordering and the quick sequential algorithm
CALIBRATION MORE OR LESS GENERALIZED

1-TRADITION: ratio, poststratification, regression

2-Calibration Principle: Goal and examples
- Examples
- Principle
- Some general properties
- A way to do that or playing with the weights

3-Generalized Calibration: Estimation knowing totals
3-1 The case of a linear link function
3-2 Properties: Bias, variance
3-3 Practical way (CALMAR II software) and applications
3-4 Some ideas on the numerical aspects
4-Standard Calibration and Applications
4-1 Choice of the instruments; choice of the link function
4-2 Determination using a distance function between the weights
4-3 Raking-ratio revisited
4-4 Calibration on inexact (unbiased) data; simultaneous calibration of several surveys

5- Two-phases Sampling and Calibration
5-1 Two phases sampling: estimator, variance and variance estimation
5-2 Auxiliary information and calibration in two phases sampling
5-3 Two phases sampling as a model for the response mechanism (an introduction)
APPLICATIONS TO NON-RESPONSE

A  -Generalities
1-Non-response is an affair of principles!!

2-Pattern of Non-response and Correction Methods
2-1 Total (unit) and partial (item) non-response
2-2 Goals of the corrections

3-Basic methods:weighting or imputing?
3-1 Outline of the methods
   a-Weighting
   b-Imputing
3-2 What Method for What Use?
B-Weighting for non-response and generalized calibration

1-Model

2-Examples: Parametric and non-parametric models

3- Estimation of the parameters
3-1 Incidence of the use of estimated parameters for estimation
3-2 How to estimate the parameters of the response mechanism?

4-Variance and estimating equations

5-Examples
C-Deterministic imputation, prediction and calibration

1-The estimator and the superpopulation model

2-Derivation from a superpopulation model

3-Estimating equations for parameters and variance

4-Choice of the instruments
4-1 Standard considerations
4-2 Trick avoiding to compute explicitly the response weights

5-Examples: ratio imputation, regression imputation

6-Summary
D-Random imputation and balanced sampling

0-Imputation ‘Weighting Like’

1-Generalities
1-1 Some new problems
1-2 Quick inventory of methods

2-Formalization and Estimation Theory
2-1 The extra-variance
2-2 Exercise (or example!)
2-3 General recipes for the extra variance
3-Applications, Examples and Concluding Remarks
3-1- Imputation of a (0-1) variable and balanced sampling
3-2- Qualitative variable and balanced sampling of cells
   REMARK: Is random imputation really mandatory?
3-3- Quantitative variable and balanced sampling of residuals
   1- Gaussian model.
   2- I like the following technique
   3- Use the Cube.
   REMARK: Is random imputation really mandatory?
3-4- Summary and final remarks
CALIBRATION, BALANCED SAMPLING WITH APPLICATIONS TO NON-RESPONSE

Jean-Claude DEVILLE
ENSAI/Crest
Laboratoire de Statistique d’Enquête
GENERALITIES

1- Universe, label, sample and sample design, inclusion probabilities

$U$ is the population of interest or Universe, $s$ is an arbitrary subset of $U$ called sample. Units of $U$ are identified by a label $k, l, ...$ as in a computer file. Conventionally those label will be considered as running from 1 to $N$, the population size.

A sample design is a probability law $p(s)$ on the set of samples: $0 \leq p(s) \leq 1$, $\sum p(s) = 1$

$S = \{s; p(s) > 0\}$ is the support of $p$; Example: fixed size sample: $\text{card}(s) = n$ fixed.

Define the (first order) inclusion probabilities: $\pi_k = \Pr(k \in s) = \sum_s p(s) \ 1(k \in s) = E_p(I_k(s))$

and the second order i.proba: $\pi_{kl} = \Pr(k \text{ and } l \in s) = \sum_s p(s) \ 1(k \in s)1(l \in s) = E_p(I_k(s)I_l(s))$

Let $n = \text{card}(s)$ be the sample size: $\sum_{U} \pi_k = E(n) = n$ if fixed size.
2- Unbiased estimation for a total

If we search an estimator among the linear estimators, \((\sum_{s} w_k (s)y_k )\), and moreover unbiased, and moreover with \(w_k\) not depending on \(s\) (if \(>0\)), there is only one estimator due to Horvitz-Thompson (HT):

\[
\hat{Y}_{HT} = \sum_{s} \frac{1}{\pi_k} y_k = \sum_{U} \mathbf{1}(k \in s) \frac{1}{\pi_k} y_k
\]

The variance is straightforward:

\[
\text{Var}\_p (\hat{Y}_{HT}) = Q(y_U) = \sum_{k,l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l}
\]

This variance is estimated without bias by:

\[
\hat{\text{Var}}_p (\hat{Y}_{HT}) = Q_s (y_s) = \sum_{k,l \in s} (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l}
\]

NB:\quad \pi_{kk} = \pi_k

For the following, it import to have in mind that those quantities are QUADRATIC FORMS. However, except of very rare cases, it is never computed as a double sum (!), and the coeff of the quadratic form are never (almost) computed.
3-Practical Sampling

Decomposition of a sampling plan: Stratification and multi-stage designs

In practice there are ‘units sampling plan’ implemented by some algorithm (see below) and two ways (formally only one !) for decomposing a plan in smaller plans. By experience, a ‘unit sampling’ does not provide a sample greater than 10 to 100 final units!

Decomposition use a ‘classification’ (auxiliary!) variable to perform the well known:

- **Stratification:** $U_h$ are strata (partition of $U$). Sample $s_h$ drawn in $U_h$ with the design $p_h$. The stratified design is $s = \bigcup_{h} s_h$ and $p(s) = \prod_{h} p_h(s_h)$ (the $s_h$ are independent random element).

With obvious notations one has: $\hat{Y} = \sum_{h} \hat{Y}_h$, $Q(y_U) = \sum_{h} Q_h(Y_{U_h})$ and $Q_s(y_s) = \sum_{h} Q_{s_h}(Y_{s_h})$.

- **Multi-stage design:** $U_i$ are ‘primary sampling units’ (psu), a sample $s_i=\{i\}$ is drawn according to a design $p_i$ in the population of primary units. For each $i$ in $s_i$ there is some design $p_i$ (conditioned by $s_i$) producing a sample $s_i$ in $U_i$ and corresponding estimates for this sub-population. With obvious notation we have:
\[ s = \bigcup_{i \in s_I} s_i \quad \text{and} \quad p(s) = p_I(s_1) \prod_{i \in s_I} p_i(s_i | s_I) \quad \text{and} \quad \hat{Y} = \sum_{s_I} \frac{\hat{Y}_i}{\pi_i} \]

For the variance and the variance estimation, the quadratic forms are defined recursively:

\[ Q(y_U) = \text{Var}(\hat{Y}) = Q_I(\{Y_i ; i \in I\}) + E_{p_I} \left( \sum_{s_I} \frac{1}{\pi_i^2} \text{Var}_{p_i}(\hat{Y}_i) \right) \]

\[ Q_s(y_s) = \text{Var}(\hat{Y}) = Q_{s_I}(\{\hat{Y}_i ; i \in s_I\}) + \sum_{s_I} \frac{1}{\pi_i} \text{Var}_{p_i}(\hat{Y}_i) \]

Remark 1: NOT TERM TO TERM ESTIMATION;
DIFFERENT ORDER OF MAGNITUDE!

Remark 2: Stratification=Multi-stage with a ‘census design’ at the first stage.
Units sampling plans:

-Bernoulli: $N$ independent draws (lotteries) with the same probability $p(s) = f^n (1 - f)^{N-n}$.
Random sample size. HT estimator $\hat{Y} = f^{-1} \sum_s y_k$. Quadratic forms $Q(y_U) = f^{-1} (1 - f) \sum_U y_k^2$ and $Q_s(y_s) = f^{-2} (1 - f) \sum_s y_k^2$.

-Poisson: $N$ independent lotteries with probabilities $\pi_k: p(s) = \prod_{k \in s} \pi_k \prod_{k \in U - s} (1 - \pi_k)$. Random sample size. HT estimator $\hat{Y} = \sum_s y_k / \pi_k$. Quadratic forms $Q(y_U) = \sum_U \pi_k^{-1} (1 - \pi_k) y_k^2$ and $Q_s(y_s) = \sum_s \pi_k^{-2} (1 - \pi_k) y_k^2$.

REM: Very important for many purposes because it is ‘the’ sampling without information and without constraints.
Simple Random Sampling: = Bernoulli conditional to fixed size \( n \): 
\[
p(s) = \binom{N}{n}^{-1}.
\]

Quadratic forms: 
\[
Q(y_U) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \sum_U \frac{(y_k - \bar{Y})^2}{N - 1}
\]
and 
\[
Q_s(y_s) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \sum_s \frac{(y_k - \bar{y})^2}{n - 1}
\]

-Conditionnal Poisson (or near maximal entropy unequal pr. Sampling).
The most important varying probabilities sampling plan.

-Balanced sample: see later.
4-Non-linear Estimation: Plug-in Principle (or substitution)

→ Examples of non-linear indicators to be estimated

- **Basics**: Ratio, mean (especially on a domain (= subpopulation!)), correlation between two variables, coefficient of a regression (or logit, or any thing!) adjustment (model if you prefer, me not!), Gini index or other concentration measures, …

- **Formally**: The set of variable used in our problem is a vector \( z_k \) for unit \( k \) living in the \( \mathbb{R}^p \) – space. The population is viewed as a set of \( N \) points in this space, each with mass equal to 1. In other words, the population is seen as a (positive, atomic, sum of Dirac masses) measure \( M \) on this space. We are interested in estimating some functional \( T(M) \) of this measure.

→ Plug-in principle for point estimation

The sample (and the weighted estimator we use, HT or another one, with weights \( w_k(s) \)) is viewed similarly as the measure with masses \( w_k(s) \) at the point \( z_k \) (if \( k \) is in the sample) and zero elsewhere. We denote this measure by \( \hat{M} \), which is in some sense an estimator of \( M \) (in fact, if
A is a subset of \( \mathbf{R}^d \) we have \( E(\hat{M}(A)) = M(A) \). The plug-in (or substitution) principles consists in taking \( T(\hat{M}) \) as an estimator \( \hat{T}(M) \).

As it is easy to see, it is the most natural one: if \( T \) is a function of totals (most current case: ratio, means, correlations, regression coefficients), the estimator is simply the same function with the totals replaced by their estimators.

→ Linearization for variance approximation with examples

Based on asymptotic ideas not discussed here. The schema is to consider \( M \) as \( N \) iid realisations of a random vector in \( \mathbf{R}^p \). The (average) sample size \( fN \) tend to infinity. The (average) sampling fraction remain constant or tend to zero. There are quantities that we can qualify to be \( O_p(1/n) \) or \( O_p(1/n^{1/2}) \) and so on. Moreover we admit that the HT estimator, divided by the scaling factor \( N \) is convergent and that a central limit theorem with the usual \( n^{1/2} \) dilatation.

A functional \( T \) is linearizable if there exist a (NECESSARILY) uniquely defined variable \( t_k = \text{lin}T_k \) such for a convenient exponent \( \alpha \) we have:

\[
N^{-\alpha} (T(\hat{M}) - T(M)) - N^{-1} \sum_s w_s t_k = O_p(1/n)
\]

It is clear that \( \alpha = 1 \) for totals, \( = 0 \) for ratios, means, medians for instance, \( = 2 \) for ‘variances’ or double sums.
Examples: - If $T$ is the total, $T(y)=Y=\sum_U y_k \Rightarrow \text{lin}T_k = y_k$.
- Functions of totals: $T=F(X,Y,\ldots) \Rightarrow \text{lin}T_k = \frac{\partial F}{\partial X} x_k + \frac{\partial F}{\partial Y} y_k + \ldots$

For the ratio $R=Y/X$, $\text{lin}R_k = 1/X \ (y_k-Rx_k)$

- Implicit function of the form $\sum l_k(B) = 0$ (where the $l_k$ are $\mathbb{R}^p \rightarrow \mathbb{R}^p$):
  \[ \text{lin}B_k = - (\sum_{U} l_k^\prime(B))^{-1} l_k(B) \]

Regression (or logit) parameters: $l_k(B)=x_k(y_k-f(x_k'B))$ or $x_k(y_k-f(x_k'B))$ with $f=\exp/(1+\exp)$.

- We shall need of the composition of functionals: $T(M,\lambda)$ where $\lambda$ is a $q$-parameter for $T$ whose value can be given by some functional. We have that: $\text{lin}T(M,\lambda(M))=\text{lin}T(M,\lambda) + \frac{\partial T}{\partial \lambda} \text{lin}\lambda$

A particular case is the ‘inverse’ function: $T(M, \lambda)$ - $S = 0 \Rightarrow \text{lin}S_k = \text{lin}T_k + \frac{\partial T}{\partial \lambda} \text{lin}\lambda_k$.

In particular, if $S$ is a constant and the dimensions agree, we have \[ \text{lin}T_k = -\left(\frac{\partial T}{\partial \lambda}\right)^{-1} \text{lin}\lambda_k. \]

A general way to obtain the linearized variable is to use the influence function, or, equivalently the estimating equation in the population which is in fact a definition.
Linearization variance estimation

One use an estimation of the approximate variance using the linearized variance. Therefore we should use $Q_s(t_k; k \in s)$. Unfortunately, the quantity $t_k$ depend always of unknown quantities of the population that have to be replaced by estimation from the sample. If the estimators are consistent, and if their number remain finite in the asymptotic set-up, the estimator of variance is consistent.

Example: For a ratio $R=Y/X$ the linearized variable is $I/X (y_k - Rx_k)$ but variance estimation will use the ‘pseudo variable’ $\tilde{t}_k = \frac{1}{\hat{X}} (y_k - \hat{R}x_k)$.

For the ratio estimator of a total $X\hat{R}$, the linearized variable is $y_k - Rx_k$.

For variance estimation you have the choice between $y_k - \hat{R}x_k$ and $\frac{X}{\hat{X}} (y_k - \hat{R}x_k)$.
5-Representativeness à la Hajek (Sampling from a finite population, chap 16).

A strategy (plan + weights) is representative with respect of the vector of variables \( x \) if for any sample \( s \) the weights \( w \) verify \( \Sigma_s w_k x_k = X = \Sigma_U x_k \). The total \( X \) has to be known and an auxiliary information.

When this information comes from a source disconnected from the sampling frame, a way is to modify slightly the unbiased HT weights such that the equality is verified. This is calibration.

When the information \( x_k \) is available in the frame, a way is to keep the unbiased HT weights and to find some sampling algorithm such that the equality is verified. This means find a method to get a random balanced sample with given inclusion probabilities.

Properties of those strategies are evaluated using the mean square error criterion and, in some sense, conditional inference.

Those tools are very useful to deal efficiently with non-response, as we shall see (I hope so!) in the sequel.
1- **Entropy and Poisson Sampling**

We are interested in sampling plans having given inclusion probabilities $0 < \pi_k < 1$ and a fixed support $S$. Otherwise, we want that the indicator variable $I_k (=1$ if $k$ in $s$, $0$ if not) are the most independent as possible, or what is nearly the same, that the value of the $p(s)$ are have a minimum of dispersion. Without making too much philosophy a good criterion to maximize is the entropy of the design $\sum_{s \in S} - p(s) \log(p(s))$ with $p(0) \log p(0) = 0$.

The $N$ constraints are $\pi_k = \sum_S 1(k \in s) p(s)$ associated to a Lagrange multiplier $\lambda_k$.

The optimisation problem leads to $p(s) = C(S) \exp(\lambda.s)$ where appear the scalar product of the $N$-vector of the $\lambda_k$ with the vector $s$ of the ‘coordinates of $s$ ‘: $1$ if it contain $k$, $0$ otherwise. The constant $C(S)$ normalize the sum of the probabilities to one.

-Suppose $S$ is the set of all subsets of $U$. Let $\exp \lambda_k = \omega_k = \pi_k(1-\pi_k^*)$.
We get that \( p(s) = C(S) \prod_{s} \omega_k = \prod_{s} \pi_k^* \prod_{U-s} (1 - \pi_k^*) \) which means that we have obtained the Poisson sampling and that \( \pi_k^* = \pi_k \).

- If \( S \) is any family of subsets of \( U \) (fixed size for instance), we have get the Poisson sampling conditional to \( S \), and the \( \pi_k \) are the conditional inclusion probabilities, ‘knowing’ that \( s \) belong to \( S \). The \( \pi_k^* \) are now different from the \( \pi_k \).

- In particular, suppose that a \( p \)-vector variable \( x_k \) is known on the sampling frame, and that we want to perform a balanced sampling design, (fixed size is a particular case) that is verifying:

\[
\sum_{s} \frac{x_k}{\pi_k} = \sum_{s} a_k = As = X = \sum_{U} x_k \quad \text{where} \quad A = \begin{pmatrix}
\frac{x_1}{\pi_1} & \frac{x_2}{\pi_2} & \cdots & \frac{x_N}{\pi_N}
\end{pmatrix}
\]

\( S \) is the family of subsets of \( U \) verifying this set of \( p \) linear constraints (approximately!).

2-Variance, Variance Approximation and Variance Estimation and Extensions

2-1 Necessity of an approximation

- In the case of fixed size conditional Poisson sampling, it is possible to compute with a great precision the second order probabilities for variance and variance estimation. However the double sum in variance estimation is not!
Moreover, it has been seen that the estimation with a simple sum of square of an approximation of the variance is much more accurate than an unbiased estimation with a double sum of the true variance.

- In the general case, and in particular for a balanced sampling exact computation of the variance seems not to be possible.

Therefore we will find a good approximation of the variance as a sum of squares and find a nearly unbiased estimation of this quantity.

. It is then true that:

a) \( \text{Var}(\sum_{s} x_k/\pi_k) = 0 \)

b) If one can write \( y_k = B'x_k + \varepsilon_k \) for some \( B \), then \( \text{Var}(\sum_{s} y_k/\pi_k) = \text{Var}(\sum_{s} \varepsilon_k/\pi_k) \) with possible consequences…

c) Suppose that we are in the asymptotic conditions for the (non-conditional) Poisson sampling that the couple \((\hat{Y}, \hat{X})\) suitably centred and normalized follows a \((p+1)\)-dimensional Gaussian (or Normal, if you prefer!) law. Its covariance matrix is given (up to the scale factor) by:
\[
\begin{pmatrix}
\sum_U \pi_k (1 - \pi_k) \left( \frac{y_k}{\pi_k} \right)^2 & \sum_U \pi_k (1 - \pi_k) \frac{x_k y_k}{\pi_k \pi_k} \\
\sum_U \pi_k (1 - \pi_k) \frac{y_k x_k}{\pi_k \pi_k} & \sum_U \pi_k (1 - \pi_k) \frac{x_k x_k}{\pi_k \pi_k}
\end{pmatrix}
\begin{pmatrix}
V_y \\
V_{Xy}
\end{pmatrix}
\]

If we consider that the conditional sampling provide an estimator which has the conditional variance, we can write:

\[
Var_{\text{PoissCond}}(\hat{Y}_{HT}) = Var_{\text{Poiss}}(\hat{Y} \mid \hat{X} = X) = V_y - V_{Xy}'V_x^{-1}V_{Xy} = Var_{\text{Poiss}} \left( \sum_s y'_k - B'x_k \right)
\]

where \( B = V_x^{-1}V_{Xy} \) is the regression of \( y/\pi \) on \( x/\pi \) with the weights \( \pi(1-\pi) \).

If \( x \) is one-dimensional and proportional to \( \pi \), the sampling is only of fixed size and we find:

\[
Q(y_U) = \frac{1}{1 - \sum U a_k} \sum U \pi_k (1 - \pi_k) \left( \frac{y_k}{\pi_k} - \left( \frac{y}{\pi} \right) \right)^2
\]

with \( a_k = \pi_k (1 - \pi_k) / \sum U \pi_k (1 - \pi_k) \) and \( \left( \frac{y}{\pi} \right) = \sum_U a_k \frac{y_k}{\pi_k} \).
A plug-in estimation of this approximation of the variance could be:

\[ Q_s(y_s) = \frac{1}{1 - \sum a_k^2} \sum (1 - \pi_k)(\frac{y_k}{\pi_k} - (y/\pi))^2 \]

with \( a_k = (1 - \pi_k) / \sum (1 - \pi_k) \) and \( (y/\pi) = \sum a_k \frac{y_k}{\pi_k} \).

The variance of the balanced sampling is given approximately by:

\[ Var_{PoissCond}(\hat{Y}_{HT}) = Var_{Poiss}(\hat{Y}|\hat{X} = X) = V_y - V_{xy}'V_x^{-1}V_{xy} = Var_{Poiss}(\sum s y_k - B'x_k) \]

where \( B = V_x^{-1}V_{xy} \) is the regression of \( y/\pi \) on \( x/\pi \) with the weights \( \pi(1-\pi) \).

We estimate the variance by estimating its approximation using the plug-in principle:

\[ V_{\hat{r}e} (\hat{Y}_{HT}) = Var_{Poiss}(\sum s y_k - \hat{B}'x_k) \]

where \( \hat{B} = \hat{V}_x^{-1}\hat{V}_{xy} = (\sum (1 - \pi_k) \frac{x_kx_k^T}{\pi_k^2})^{-1} \sum (1 - \pi_k) \frac{x_ky_k}{\pi_k^2} \) is the regression of \( y/\pi \) on \( x/\pi \) with the weights \( 1 - \pi \).
3-Exact or approximate balance?

A difficulty is that, for many variables, we cannot get an exact balance because simply of arithmetical difficulties. Using again an asymptotic argument, we can say that an HT estimator which is not balanced (say a Poisson) has an error which is $O_p(n^{-1/2})$. When we balance on some variable, its HT estimator becomes a $O_p(n^{-1})$.

There exists a necessary and sufficient condition for exact balance which is that each full rank matrix of $A$ has the same determinant in absolute value. It implies that $A$ is equivalent to a matrix having for entries only 1, 0 or -1 (necessary condition). A sufficient condition is the case of quota sampling on the margins of a contingency table.
4-The Cube Method

4-1 The Cube set-up
Let \( K = \{x; Ax = X\} \) the intersection of the N-cube and the ‘plane’ (affine variety in regular maths!) of the constraints. Each summit of the cube is a (possible sample) and is denoted \( s \) as usual… \( K \) is a convex polyhedron (simplex) with extremal points having a particular place in the cube: each of them lies in a p-face described by a partition \((s_0, s^*, s_1)\) of \( U=[1,N]\) with coordinates \( 0 \) in \( s_0 \), \( 1 \) in \( s_1 \) and verifying belonging to \([0,1]\) in \( s^* \) and, moreover, \( \text{card}(s^*)<p+1 \). To get a sample, we will first draw at random one of this extremal point –this is the ‘flight phase’-, and then ‘round up’ randomly the coordinates in \( s^* \) to get a sample –this is the ‘landing phase’-. As we can see easily, every extremal point verify the constraints, and the ‘chosen’ sample \( s \) will verify it approximately in this sense that we will have \( |A s - X| \leq p \text{Max}_U |a_k| \). The problem is ‘exact’ if \( s^* = \phi \) for all extremal points. It is verified in particular cases such as fixed size sampling (only one constraint!) or marginal quota sampling. Our goal is to get a random sample verifying a set of given inclusion proba \( \pi_k \) (seen as a vector \( \pi \) which lies in the interior of \( K \) by the definition of the constraints) and verifying the given constraints (at least approximately in the preceding sense).
4-2 Flight (iterative) phase
Start from $\pi$ and find a direction $d$ in $K$ verifying $Ad=0$. The line $\pi + t d$ cut the frontier of $K$ at two points $\pi + \alpha d$ and $\pi - \beta d$, $\alpha$ and $\beta$ positive. Draw $\pi^{(t)}$ between those two points with proba $\beta/(\alpha+\beta)$ for the first one and $\alpha/(\alpha+\beta)$ for the second. It defines one coordinate $k_1$ to put to 0 or to 1 and $\pi^{(t)}$ lives in a $N-1$ face of the cube, that is in a $N-1$-cube. Now, iterate by choosing a direction in this cube verifying $As=0$. After at most $N-p$ iteration we have reach an extremal point of $K$ and the algorithm cannot go further. Each step $t$ gives a random result $\pi^{(t)}$ such that $E(\pi^{(t+1)}|\pi^{(t)})=\pi^{(t)}=\pi$. So, at the last step, we have get a random extremal point $s^*$ of $K$ verifying $E(s^*)=\pi$.

4-2 Landing (round-up) phase
If $s^*=(s_0,s^*,s_1)$ (at least symbolically!) we have to find an unbiased rounding up of $s^*$, that is a sampling plan $p_{s^*}$ on this set verifying $E_{p_{s^*}}(s_k)=s^*_k$ (with self-speaking notations). This can be done in many way (Poisson, relax some constraint and continue the cube-algorithm, find a minimum variance plan for $\sum a_k^2$, ...).
5-Variance and Variance Estimation

How to get maximum entropy?
So far, we have no indication for the variance of the plan we have obtained except in the maximum entropy case (and of Gaussian approximation). An intuitive idea is that we will get nearly maximum entropy if we draw at each step the direction \( d \) ‘the most at random’. Another way is to define a sequential procedure and to perform it on a randomly sorted file.

Ordering and the quick sequential algorithm
This procedure is the following:

At the first step: \( \pi^{(1)} = \pi \), \( s_0 \) and \( s_1 \) are empty. The \( px(p+1) \) working matrix is \( A^\pi = \{a_k, k=1,p+1\} \). We suppose, for simplification, it is always of rank \( p \).

There is a vector \( b \) in \( R^{p+1} \) verifying \( A^\pi b = 0 \). Therefore the direction \( d = (b,0) \) verify \( Ad = 0 \) and becomes our candidate for the iteration process of the flight phase:

‘The line \( \pi^+ t d \) cut the frontier of \( K \) at two points \( \pi^+ \alpha d \) and \( \pi^- \beta d \), \( \alpha \) and \( \beta \) positive.
Note that only the coordinates present in \( A^\pi \) are changed. Draw \( \pi^{(1)} \) between those two points with proba \( \beta/(\alpha+\beta) \) for the first one and \( \alpha/(\alpha+\beta) \) for the second. It defines one coordinate \( k_1 \) to put to 0 or to 1 and \( \pi^{(1)} \) lives in a N-1 face of the cube, that is in a N-1-cube’.
The coordinate $k_1$ is stored in $s_0$ or $s_1$ according to the result and $a_{k1}$ is replaced by $a_{p+2}$ in $A^\wedge$. The rest of the algorithm is straightforward ($a_{p+t}$ enters in $A^\wedge$ at the $t^{th}$ step).

This procedure has two advantages. We have already pointed out that with the use, of a random ordering of the file, it provides us (we think so!) a nearly Max entropy procedure, justifying the above variance approximation.

The second advantage is practical: there is no limitation to the size of the file on which the cube method can be applied. All the calculations at one step, resume essentially to the resolution of a small set of linear equations.
# CALIBRATION MORE OR LESS GENERALIZED

## 1-TRADITION:

<table>
<thead>
<tr>
<th></th>
<th>Auxiliary information</th>
<th>$w_k$</th>
<th>Prediction error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ratio</strong></td>
<td>$X = \sum_U x_k$</td>
<td>$\frac{X}{\hat{X}}d_k$</td>
<td>$y_k - Rx_k$</td>
</tr>
<tr>
<td><strong>Postsatification</strong></td>
<td>$X = \begin{pmatrix} N_h \vdots \end{pmatrix} = \sum_U \begin{pmatrix} 1(k \in U_h) \vdots \end{pmatrix}$</td>
<td>$1(k \in U_h)\frac{N_h}{\hat{N}_h}$</td>
<td>$1(k \in U_h)(y_k - \bar{Y}_h)$</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>$X = \sum_U x_k$</td>
<td>$d_k(1 + (X - \hat{X})'T_s^{-1}q_k x_k)$</td>
<td>$y_k - B'x_k$</td>
</tr>
<tr>
<td><strong>Raking ratio</strong></td>
<td>$N_{i+}(i = 1 \text{ to } I)$ $N_{+j}(j = 1 \text{ to } J)$</td>
<td>$d_k A_i B_j \text{ if } k \in s_{ij}$</td>
<td>$y_k-a_i+b_j$</td>
</tr>
</tbody>
</table>

(ANOVA model)
2-Calibration Principle

Goal and examples

The set of variable used in our problem is a vector $z_k$ for unit $k$ living in the $\mathbb{R}^p$ -space. The population is viewed as a set of $N$ points in this space, each with mass equal to 1. In other words, the population is seen as a (positive, atomic, sum of Dirac masses) measure $M$ on this space. We are interested in estimating some functional $T(M)$ of this measure, and, moreover, we assume that we know exactly, by an external source, the value of an $\mathbb{R}^p$-valued functional $S(M)$. Using the plug-in principle, we get an estimate of this functional, namely $\hat{S}(\hat{M})$ (recall that $\hat{M}$ is the punctual measure with masses $w_k$ for units in the sample).

→ Examples:

- Estimation of a ratio $R=Y/X$ knowing a set of other ratios,
- Estimation of the median of a distribution knowing the mean, standard deviation, skewness,…,
- Known totals, for example the margins of some contingency table.

The known elements are $S(M)$, $\hat{S} = S(\hat{M})$ and $\hat{T} = T(\hat{M})$. We seek for an improved estimator $\hat{T}_{cal}$ for $T$.

An intuitive idea for the median/mean problem: translate (or scale) the whole empirical distribution…
\textbf{Principle}

Accept to loose $p$ degrees of freedom in the data, using an ‘adjustment parameter’ $\lambda \in \mathbb{R}^p$. Define (how? Don’t be in a hurry!) a family of functionals $S(M, \lambda)$ verifying $S(M, 0) = S(M)$ for any measure on which the functionals are defined. In particular we are able to work with the $S(\hat{M}, \lambda)$. If ‘everything is regular’, $S(M, \lambda)$ maps, when varying $\lambda$ a neighbourhood of 0 onto a neighbourhood of $S$. At the same time $S(\hat{M}, \lambda)$ maps, when varying $\lambda$ a neighbourhood of 0 onto a neighbourhood of $\hat{S}$. If $\hat{S}$ is not to far from $S$, the equation(s):

$$S = S(M, 0) = S(\hat{M}, \lambda) \quad \text{(CALIBRATION EQUATION(s))}$$

has a uniquely defined solution, say $\hat{\lambda}$ (see picture).

Suppose we are able, for instance using a similar construction, to define a set $T(M, \lambda)$ ‘around’ $T$, the (very generalized!) \textbf{calibration estimator} will be:

$$\hat{T}_{cal} = T(\hat{M}, \hat{\lambda})$$
\[ S = S(0) = \hat{S}(\hat{\lambda}) \]

\[ \hat{T}(\hat{\lambda}) \]

\[ \hat{S}(0) = \hat{S} \]
Example:

\[ S_\lambda = (I_p + \text{diag}(\lambda)) S \], and \[ T_\lambda(1 + b^T \lambda) \] for some given (how?) vector \( b \). The calibration equations are trivial \( S = (I_p + \text{diag}(\lambda)) \hat{S} \) and give: \( \tilde{\lambda} = \text{diag}(\hat{S})^{-1}(S - \hat{S}) \).

The calibrated estimator becomes \( \hat{T}_{cal} = \hat{T}(1 + \sum_i b_i \frac{\hat{S}_i - S_i}{\hat{S}_i}) = \hat{T}(1 + \sum_i b_i) - \sum_i b_i \frac{S_i \hat{T}}{\hat{S}_i} \), a formula full of good sense!! In particular if \( S \) is one dimensional and we take \( b = -1 \), we get \( \hat{T}_{cal} = \frac{S_i \hat{T}}{\hat{S}_i} \), a kind of ratio estimator which seems to be very efficient in some cases.
Some general properties:

- Asymptotically unbiased (because \( \hat{\lambda} \) can be seen as a substitution estimator –of \( \theta \) but it does not matter!)

- Variance can be computed by linearization. Let \( s_k = \text{lin} S_k \) and \( t_k = \text{lin} T_k \); \( \text{lin} \lambda_k = - \left( \frac{\partial S}{\partial \lambda} \right)^{-1} s_k \)

and \( \hat{T}(\hat{\lambda}) \) (with a slight abuse of notation) has for linearized variable \( t_k - \frac{\partial T}{\partial \lambda} \left( \frac{\partial S}{\partial \lambda} \right)^{-1} s_k \)

(where the partial derivatives of \( T \) are written as a line vector).

Following of the example just above:

\[
\frac{\partial S}{\partial \lambda} = \text{diag}(S), \quad \frac{\partial T}{\partial \lambda} = b', \quad \text{and the linearized variable is} \quad t_k - b' \text{diag} \left( \frac{T}{S_i} \right) s_k.
\]

For the case \( p=1 \) we get the traditional(?) ratio \( t_k - T/S s_k \).

A way to do that

A way to construct families of functionals is to play with the weights. Starting from the measure \( M \), with arbitrary masses denoted by \( w_k \) (\( I \) for the population, \( d_k I(k \ in \ s) \) for the sample), we define the (pointual) measures \( M_{\lambda} \) with masses \( w_k F_k(\lambda) \) where the \( F_k \) are, for the moment,
arbitrary regular functions $\mathbf{R}^p \to \mathbf{R}$ verifying $F_k(0) = 1$ and having a limited development having the form: 

$$F_k(\lambda) = 1 + z'_k \lambda + O(\|\lambda\|^2).$$

The gradient at 0 of the $F_k$ is a vector of variables $z_k$ that we will call (for reasons which will appear soon!) the instruments (or instrumental variables). If you have metaphysical anxiety with the $F_k$, consider that you always can take a linear, or trust in the following!

The functional $S$ and $T$ being supposed defined for all (at least) punctual measures, quantities like $S_\lambda = S(M_\lambda)$, $\hat{S}(\lambda) = S(\hat{M}_\lambda)$ are perfectly defined, and the calibrations equations write very simply (it’s only an appearance!) $S = \hat{S}(\hat{\lambda})$.

Let us give an example: $T=Y/X$ (ratio) is to be estimated, and $s_k = y_k / x_k$ is observed on the sample and available on the frame.

One can build a weighted estimator with the calibration function: (sample $s<$)

$$F_k(\lambda) = 1 - \lambda \quad \text{if} \quad s_k < \bar{S}$$

and (sample $s>$)

$$F_k(\lambda) = 1 + \lambda \quad \text{if} \quad s_k \geq \bar{S}$$
The calibration equation is: (instrument: $1(s_k \geq \bar{S}) - 1(s_k < \bar{S})$)

$$S = \sum_{s < \bar{s}} w_k s_k (1 - \lambda) + \sum_{s \geq \bar{s}} w_k s_k (1 + \lambda) = \sum U S_k$$

And finally, the calibrated estimator is:

$$\hat{T}(\hat{\lambda}) = \frac{\hat{Y} + \hat{\lambda} (\hat{Y}_{\geq} - \hat{Y}_{<})}{\hat{X} + \hat{\lambda} (\hat{X}_{\geq} - \hat{X}_{<})}$$

**Variance:** Applying the results of 1-2, we have: \[
\frac{\partial S}{\partial \lambda} |_{\lambda=0} = \sum U s_k z' \text{ where } z_k = \frac{\partial F_k}{\partial \lambda} |_{\lambda=0}
\]

Therefore $\text{lin} \lambda_k = l_k = -(\sum U s z')^{-1} s_k$, and as $\frac{\partial T}{\partial \lambda} |_{\lambda=0} = \sum U t_k z_k$, the linearized for $\hat{T}(\hat{\lambda})$ is $t_k - B's_k$ where $B = (\sum U s_k z_k')^{-1} \sum U t_k z_k$.

As it can be seen, $B$ is the instrumental regression of $t$ on $s$ using $z$ as instruments.
3-Generalized Calibration

Estimation of a total knowing a vector of totals
It is easier if $T$ is a total and $S = X = \sum_{U} x_k$ a vector of totals.

Remarks: $S_i$ can be a ratio $R = Y/X$ because knowing $R$ is the same as knowing that the total of
the (linearized! It is not completely by chance!) variable $y_k - Rx_k$ is equal to 0. It may also be a
quantile $m_{\alpha}$, because it means that $\sum_{x_k < m_{\alpha}} 1$ is equal to $\alpha N = \alpha \sum_{U} 1$, a variation of the ratio.

The calibration equations are:

$$X = \sum_{s} d_k x_k F_k(\lambda)$$
3-1 The case of a linear link function

We take: \( F_k(\lambda) = 1 + z_k' \lambda \) which leads to weight having the form: \( w_k = d_k (1 + z_k' \lambda) \).

This is an important theoretical case, but NOT the most practically used!

We get: \( \hat{\lambda} = (T_{szx})^{-1} (X - \hat{X}) \) with \( T_{szx} = \sum_s d_k z_k x_k' \). We suppose \( T_{szx} \) of full rank.

\[
\hat{Y}_{cal} = \sum_s d_k y_k (1 + (X - \hat{X})' T_{szx}^{-1} z_k)
\]

Therefore:
\[
= \hat{Y}_{HT} + (X - \hat{X})' T_{szx}^{-1} \sum_s d_k z_k y_k
\]
\[
= \hat{Y}_{HT} + (X - \hat{X})' \hat{B}
\]

With \( \hat{B} = T_{szx}^{-1} \sum_s d_k z_k y_k \) solution of the normal equations \( \sum_s d_k z_k (y_k - x_k' \hat{B}) = 0 \) of the the instrumental regression of \( y \) on \( x \) using \( z \) as instruments.

**Variance is obtained using the residual trick.** You have just to use the residuals of the instrumental regression \( e_k = y_k - x_k' B \) and \( Var(\hat{Y}_{cal}) = Q(e_U) \). Variance estimation follows on the same line: \( \hat{Var}(\hat{Y}_{cal}) = Q_s(\hat{e}_k; k \in s) \) with \( \hat{e}_k = y_k - x_k' \hat{B} \).

**Instruments need to be known only on the sample (they are not auxiliary information).**
Examples:

1-Ratio estimator: $X = \sum_{s} d_k x_k (1 + z_k \lambda)$ but $z_k = 1$ and $\hat{B} = \frac{\hat{Y}}{\hat{X}} = \hat{R}$. Residuals: $y_k - Rx_k$.

2-Weighted regression: $z_k = q_k x_k$

3-Optimal (Montanari) estimator:

$$z_k = \sum_{\ell \in s} \Delta_{k\ell} x_\ell \text{ with } \Delta_{k\ell} = \frac{\pi_{k\ell}}{\pi_k \pi_\ell} - 1; \text{For stratification one finds: } z_k = (x_k - \bar{X}_h) \frac{N_h^2}{n_h} (1 - \frac{n_h - 1}{N_h - 1}).$$

4-A nonlinear example:
Let be $u_k > 0$, $\lambda = (a, b, c)$ and $F_k(\lambda) = a + \exp(bu_k) u_k^c$. We have: $F_k(0) = 1$

$$\frac{\partial F_k}{\partial a} = 1$$

and:

$$\frac{\partial F_k}{\partial b} = u_k \exp(bu_k) u_k^c \Rightarrow \frac{\partial F_k}{\partial b}(0) = u_k$$

$$\frac{\partial F_k}{\partial c} = \log u_k \exp(bu_k) u_k^c \Rightarrow \frac{\partial F_k}{\partial c}(0) = \log u_k$$
3-2 Properties: Bias, variance

They can be derived from the general properties in 1-3, in particular the asymptotic unbiasedness and the variance obtained by the residual trick. However a more pedestrian way can be adopted with the hypotheses: $\sum_{U} z_k x'_k$ converge to a full rank matrix, $\text{Max}(\|z_k, x_k, F'_k\|)$ remain bounded.

The (generalized) calibration estimator is convergent, asymptotically (design) unbiased. Its variance is obtained with the residual trick using the residuals of the instrumental regression of $y$ on $x$ with the instruments $z_k = \frac{\partial F_k}{\partial \lambda} |_{\lambda=0}$. As a consequence, all calibrated estimators having the same $z_k$ have the same (asymptotic) variance.

One question of interest could be to find the best instruments. In fact the question has an apparently clear answer: see the example above with the Montanari estimator. In practice, it is not so clear because the $z_k$ are linear combination of the $x_l (l \text{ in } U)$ and must be estimated. The other question is the way calibration works when there is non-response.
3-3 Practical way (CALMAR II software) and applications
Generalized calibration is available in the CALMAR II software on the web site of INSEE (I believe!). Its major interest, as it will be seen later, is to perform at the same time a correction of the bias of response and a reduction of variance by taking into account the auxiliary information.

3-4 Some ideas on the numerical aspects
The numerical solving of the calibration equation is instructive. We have to solve (in the generalized linear form):

$$\Phi_s(\lambda) = \sum_s d_k x_k F(z_k \lambda) = X$$

Newton method gives:

$$\lambda_{n+1} = \lambda_n + \Phi'_s(\lambda_n)^{-1}[X - \Phi_s(\lambda_n)]$$

With a first iteration: $$\lambda_0 = 0$$, we get: $$\lambda_1 = T_{szx}^{-1}(X - \hat{X})$$, which means that the first iteration is just the regression estimator. As Newton method converge at geometrical speed, for a given set of instruments, the value of $$\lambda$$ will not vary to much and the choice of the $$F$$ function(s) has essentially the interest of ‘controlling’ the tail of the distribution of the weights.
4-Standard Calibration and Applications

4-1 Choice of the instruments; choice of the link function

The choice of the instrument is nearly equivalent to the choice of a regression.

In Standard Calibration (The first version of the CALMAR software), the choice is \( z_k = x_k \) or \( z_k = q_k x_k \) the \( q \) being weights for the regression induced by some modelisation ideas involving heteroschedastic independent errors (EX: ratio estimator; but what happens if \( x_k = 0 \) ?). Moreover, there is only a unique function \( F \), that is the weights have the form \( d_k F(q_k x_k' \lambda) \). How to choose \( F \)? If you fake it linear \( F(u) = 1 + u \), it is well known and quite straightforward that you obtain the GREG (Generalized REGression estimator).
Determination using a distance function between the weights

We seek for : New weights $w_k$ such that :
- They verify the constraints $\sum w_k x_k = X$
- They are near of the initial weights, hoping they will inherit some of their properties.

**Proximity** : ‘distances‘  $G_k(w,d)$
- $\geq 0$
- $G_k(d,d) = 0$
- $G_k(.,d)$ is convex two times continuously derivable

$g_k(.,d) = \frac{\partial}{\partial w} G$ is continuous increasing, $g_k(d,d) = 0$

$g_k(.,d)^{-1}$ exist (at least in a neighborhood of d) and may written $d_k F_k(.)$
\forall s: \text{Min } \sum_s G_k(w_k, d_k) \text{ with } \sum_s w_k x_k = X

Using a vector \( \lambda \) of Lagrange multipliers:

\forall k \in s: g_k(w_k, d_k) - x_k' \lambda = 0 \quad \Rightarrow w_k = d_k F_k(x_k' \lambda)

which leads to the calibration equations

\[ \sum_s d_k x_k F_k(x_k' \lambda) = X \]
In practice (CALMAR !) we restrain ourselves to: \( G_k(w,d) = \frac{d}{q_k} G\left(\frac{w}{d}\right) \), \( G \) convex, two times continuously derivable, verifying \( G(1) = 0, \ G'(1) = 0, \ G''(1) = 1 \)

\[ \Rightarrow g_k(w,d) = \frac{1}{q_k} G'\left(\frac{w}{d}\right) \quad \text{and} \quad w_k = d_k F(q_k x_k \lambda). \]

Where \( F \) is the reciprocal function of \( G' \).

<table>
<thead>
<tr>
<th>Distance ((x = \frac{w}{d}))</th>
<th>Calibration on Margins</th>
<th>F(u) Calibration Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{2}(x-1)^2 \Rightarrow \sum_s \frac{(w-d)^2}{2d} )</td>
<td>( \frac{(\hat{N}_w - \hat{N})^2}{\hat{N}} )</td>
<td>( 1+u ) ( \text{linear} )</td>
</tr>
<tr>
<td>xLogx – x +1</td>
<td>( \hat{N}^w \log \frac{\hat{N}^w}{\hat{N}} + (\hat{N} - \hat{N}^w) )</td>
<td>exp (u) Raking-Ratio</td>
</tr>
<tr>
<td>Expression</td>
<td>Description</td>
<td>Condition</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>----------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>((x - L) \log \frac{x - L}{1 - L} + (U - x) \log \frac{U - x}{U - 1})</td>
<td>Si (x \in [L, U])</td>
<td></td>
</tr>
<tr>
<td>( \frac{(x - 1)^2}{2} ) si (x \in [L, U])</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- (\log x + x - 1)</td>
<td>(\hat{N} \log \frac{\hat{N}_w}{\hat{N}} + (\hat{N}_w - \hat{N}))</td>
<td></td>
</tr>
<tr>
<td>(x + \frac{1}{2} - 2)</td>
<td>(\frac{(\hat{N}_w - \hat{N})^2}{\hat{N}_w})</td>
<td></td>
</tr>
<tr>
<td>((\sqrt{x} - 1)^2 \Rightarrow (\sqrt{w} - \sqrt{d})^2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\frac{L(U - 1) + U(1 - L) \exp Au}{(U - 1) + (1 - L) \exp Au})</td>
<td>(A = \frac{U - L}{(U - 1)(1 - L)}) (log it)</td>
<td></td>
</tr>
<tr>
<td>(1 + u) si (x \in [L, U])</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(truncated linear)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((\chi^2))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((\text{Hellinger}))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\frac{1 + \alpha u}{1 + \alpha u}^{1/\alpha})</td>
<td>(\alpha \neq 0)</td>
<td></td>
</tr>
</tbody>
</table>
**LINEAR METHOD**: CALIBRATING EQUATIONS

\[
\sum_{s} d_k q_k x_k \left( 1 + q_k x_k' \lambda \right) = X \Rightarrow \lambda + T_s^{-1} \left( X - \hat{X} \right) \quad T_s = \sum_{s} d_k q_k x_k x_k'
\]

Weights: \[ w_k = d_k \left( 1 + \left( X - \hat{X} \right)' T_s^{-1} X \right) \]

Regression estimator !! (from which ratio, poststratification are particular cases).

**EXPONENTIAL METHOD**

You find exactly the raking-ratio method if all \( q_k = 1 \) and auxiliary information is reduced to margins of a table.

However you can add other constraints

**EXAMPLE**: Age and region structure + average income
◆ **LOGIT AND TRUNCATED LINEAR METHODS**

They allow to get weights having a bounded range, to avoid negative weights (linear method) or excessively high weights (raking-ratio).

◆ **OTHER METHODS**

Often rises numerical difficulties but may have strong theoretical or practical justifications (for example a criterion on maximization of entropy which is connected to maximum likelihood principle-empirical likelihood-).
4-3 \rightarrow \textbf{Raking-ratio revisited}

We have already seen that the raking-ratio weights are solution of ‘calibration equations:

\[
\sum_{s} d_k \begin{bmatrix} I_k \\ J_k \end{bmatrix} \exp\left( I_k \alpha + J_k \beta \right) = \sum_{U} \begin{bmatrix} I_k \\ J_k \end{bmatrix}
\]

We get a standard calibration estimator with \( F = \exp \), \( q_k = 1 \), and \( X = (N_{i+}, N_{+j}) \).

\rightarrow \text{The variance of the raking-ratio estimator is obtained by the residual trick. You have to take as new variables the residuals of a ANOVA model :}

\rightarrow \[ y_k = \bar{Y} + A_i + B_j + \text{residual} \]

4-4 \rightarrow \textbf{Calibration on inexact (unbiased) data ; simultaneous calibration of several surveys}

Outline: The inexact data are seen as a concurrent unbiased estimator \( X^* \) of \( X \) with a known variance. Compute the best unbiased linear estimator \( X^{**} \) derived from \( \hat{X} \) and \( X^* \). Calibrate on \( X^{**} \).

A typical case occurs when the auxiliary data come from another survey. The method can be seen as a simultaneous calibration of the two surveys. For more than two surveys, the method is
nearly the same: compute the best linear estimator $X^*\cdots^*$ derived from all the surveys and calibrate on the result.

NB: Exact data have zero variance, missing ones infinite. With this convention exact and missing information have the same status!

5- Two-phases Sampling and Calibration
5-1 Two phases sampling: estimator, variance and variance estimation

Take a sample $r$ in $s$ considered as a population (Neyman 1938!) giving a (cheap) auxiliary information for stratification or regression estimation: $U \rightarrow s \rightarrow r$ 

$p(s) \rightarrow q(r/s)$
$$\pi_k = \sum_{s \in k} p(s) \quad \text{InclusionProbabilities} \quad \pi_{kl} \quad \text{second order proba}$$

$$P_{k/s} = \sum_{r \in (k, s)} q(r/s) \quad \text{Second phase Probabilities} \quad \text{(notation : we drop the /s)}$$

$$P_{kl/s} = \sum_{r \in (k, \ell)/s} q(r/s) \quad \text{Second phase second order Probabilities}$$

Basic estimator (for a total) ( expansion estimator):

$$\hat{Y}_{exp} = \sum_{r} \frac{y_k}{\pi_k P_{k/s}}$$

Unbiasedness :

$$E_q \left( \hat{Y}_{exp} \mid s \right) = \sum_{s} \frac{y_k}{\pi_k} = \hat{Y}_{HT} \quad \text{et} \quad E_p \hat{Y}_{HT} = Y$$

Variance:

$$Var \hat{Y}_{exp} = Var E \left( \hat{Y}_{exp} \mid s \right) + E Var \left( \hat{Y}_{exp} \mid s \right) = Var \hat{Y}_{HT} + E \left( \sum_{s} \sum_{\ell} \Delta_{k\ell} y_k y_\ell \frac{\pi_k}{\pi_\ell} \right) \Rightarrow Q(y_U) = Q_1 + Q_2$$

Variance estimation :

$$\sum_{s} \sum_{\ell} \Delta_{k\ell} y_k y_\ell + \sum_{s} \sum_{\ell} \Delta_{k\ell}^2 y_k y_\ell \frac{\pi_k}{\pi_\ell}$$
estimated by:

\[ \sum \sum \frac{\Delta^{(1)}_{k\ell}}{P_{k\ell}} y_k y_\ell + \sum \sum \frac{\Delta^{(2)}_{k\ell}}{\pi_k \pi_\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} = Q_1r + Q_2r \]

We make the assumption that this estimation does not make a problem !!!!

5-2 Auxiliary information and calibration in two phases sampling

- You can calibrate the two phases expansion estimator exactly like the HT estimator if you have an information on \( U \). You get the calibration equations:

\[ X = \sum \frac{x_k}{\pi_k P_k} g_k \]

Let be \( y_k = x'_k B + e_k \Rightarrow \hat{Y}^w = X'B + \sum g_k \frac{e_k}{\pi_k P_k} \) : Variance and variance estimation can be obtained with the residuals trick:

\[ \text{Var} (\hat{Y}_{\text{exp}}) = Q_1(e_U). \]

- Suppose that \( X \) is not known and that \( \hat{X}_{HT} \) is known (the variables \( x_k \) are measured in \( s \)), you can calibrate on \( \hat{X}_{HT} \) \( \Rightarrow \) Variance = \( \text{Var} \hat{Y}_{HT} + E_p \text{Var}_q (\hat{E}|s) = Q_1(y_U) + Q_2(e_U). \)
- If you know $X$ and $\hat{X}_{HT}$, you can calibrate in two steps or calibrate directly on $X$. You don’t get necessarily the same weights. However, if you have in mind only the problems about variance, the two methods give equivalent results.

- **In the general case** an auxiliary vector $X_0$ known in $U$, another one $\hat{X}_{1,HT}$ is known from $s$. Recipe: calibrate on the most informative data. That means that we have to use the calibration equations:

$$\begin{pmatrix} X_0 \\ \hat{X}_1 \end{pmatrix} = \sum_r d_k P_k^{-1} \begin{pmatrix} x_{0k} \\ x_{1k} \end{pmatrix} F_k(\lambda) \quad \text{with eventually} \quad F_k(\lambda) \cong 1 + z_k' \lambda$$

The variance is given by $Q_1(e_s^U) + Q_2(e_U)$ where $e = y - \tilde{B}_0 x_0 - \tilde{B}_1 x_1 = \text{Residuals}(y/(x_0, x_1)$ using $z)$ and $e_s = y - \tilde{B}_0 x_0 = \text{Residuals}(y/x_0$ using $(x_1, z^\perp)$. ($z^\perp$ is the orthogonal supplement!). (see picture). The instrument are therefore given by: $x^*_{0k} = T_{rx_0z} T_{rzz}^{-1} z_k$
5-3 Two phases sampling as a model for the response mechanism (an introduction)

Technically, if you know perfectly how works the response mechanism, you are working with a two phases survey (or three if your initial survey was designed in two phases!): The first phase was perfectly controlled and the second is only observed and the parameters of the sampling mechanism are unknown.
APPLICATIONS TO NON-RESPONSE

Generalities:

1-Non-response is an affair of principles!!

1-1 First principle: TRY TO AVOID IT! (practical aspects!)

1-2 Second principle: Observe, analyze, correct
- OBSERVE non-response and don’t forget that it is a true problem (data collecting report, participate yourself to data collection,..)

- ANALYZE non-response with the same attention as any statistical fact.

- TRY TO CORRECT the bias which is an $O_p(1)$!!

- DOING NOTHING IS DOING SOMETHING!!
You need at least for a correction for totals because your sample is not as big as you expected.

$\Rightarrow$ model the response mechanism $q(r \mid s)$.

Total = homogeneous of degree 1 versus ‘deformations’ homogeneous of degree 0.
Three steps

In practice, the response mechanism is often modelized by Poisson or conditional Poisson sampling. The response probabilities are taken as functions of some vector of parameters $\beta$ (in the parametric approach), often as a linear generalized model (log-linear or logit):

$$\frac{1}{P_k} = F_k(\beta) = F(z_k, \beta) = F(z_k \beta).$$

- **First step**: select the good variables $z$. (and not the most easy ones!- if possible..-).

**IMPORTANT WARNING**: Among the $z$ it is mandatory to include the ‘free of charge constant variable’ $1$, eventually as a known linear combination of $z$. It allows to take into account the overall level of non-response (for estimating totals and other statistics homogenous of degree 1), the other variables taking in charge the structure of the sample (essential for ratios and other statistics homogenous of degree 0). For the rest of those notes this will be the case. Logit response models give a very convincing illustration of this fact-see later.

- **Second step**: Formulate a model and a method for correction (there is a difference!)

  **Remark**: it is often convenient to simplify the model (enlarging bias) if variance decreases. Sometimes also, simplification makes the correction more robust.
-Third step: Incorporate the correction (for bias if your model is ‘true’) in variance estimation. Example: A nice logit model can predict some unacceptable response probabilities and will be replaced by a cell model (homogenous response classes in the yellow book- Särndal, Swennson, Wretman).

2-Pattern of Non-response and Correction Methods
2-1 Total (unit) and partial (item) non-response
Individual information can be present at different levels: Population $U$ (sampling frame), sample $s$ (information easy to collect, involving not necessary a direct contact with the sampled unit), the set $r$ of respondent (for which a minimal amount of information has been collected). The set $o = s - r$ is the set of total(or unit) non-response. For each variable $v$, there is a set $r_v = r - o_v$ of partial response specific to the variable (see table). Partial non-response sets can be articulated in all possible fashions.
<table>
<thead>
<tr>
<th>Frame of the population $U$</th>
<th>Sample $s$</th>
<th>Respondant $r$</th>
<th>Total non-response $o$</th>
<th>U-s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


2-2 Goals of the corrections

One variable: total, mean, mean on a domain, distribution function, quantiles, concentration measures, in fact all derived from the distribution function.

Two or more variables: covariance, regressions, principal components, two way classifications, ...

In fact there are linear stat (total, domain total,?) and non linear stat (bound to the df).
3-Basic methods: weighting or imputing?

3-1 Outline of the methods

**a-Weighting:** new weights to compensate non-response and taking into account the response probabilities.

**b-Imputing:** replace the missing information by a ‘good’ (= acceptable?) value.

Two ways: **deterministic imputation** versus **random imputation**.

**Based on statistical models estimations.** Parametric or non-parametric. More explicitly:

Weighting: \(1 / \Pr(k \in r | s) = (para) g_k(\beta) = (event) g(z_k \beta) = (non \ para) g(z_k)\) with \(g\) to be estimated.

Deterministic imputation: -Find a predictor \(\hat{y}_k = f(x_k)\) (non - para) = \(f(x_k, \theta)\) (para)

Don’t forget, in the two cases, that the estimations are made on \(r\) (\(\Rightarrow \hat{y}_k = f(x_k, r)!!\)).
Random imputation.-Find a probability law for $y_k$ that is $\widetilde{y}_k = \hat{y}_k + \varepsilon_k$. The law can be researched parametrically (normal law for instance), non-parametrically (the empirical law of some set of respondent for instance), or semi-parametrically (often, the predictor is parametrically estimated but the law of the residuals is non parametric). In all cases, the estimations are based on the set of respondents.

3-2 What Method for What Use?

Here is a table (in great part justified by the following) giving the performance of the different methods for the different problems. In fact, if we have in mind THE ‘textbook’ variable of interest, we have not to take into account the bi-multi-dimensional problems.

It is clear that weighting has a clear superiority because everything is OK!

Imputation by prediction suffers from the impossibility of tackling estimation of non-linear quantities (ex: imputation by the mean), random imputation using a proba law provide an unnecessary extra variance (ex: hot-deck= drawing ‘at random’ some ‘donor’ among the respondents). Moreover, variance estimation is quite natural and easy with weighting, but becomes complicated and misleading with imputed data. If the pattern of non-response is the same for a set of variables, the same remarks apply to multivariate analysis (dependence between variable are likely to be respected by weighting).

Therefore, weighting is the best way to deal with complete non-response.
<table>
<thead>
<tr>
<th>Weighting</th>
<th>Total Mean, domain</th>
<th>Distribution function</th>
<th>Bidimensional</th>
<th>Multidimensional</th>
<th>Variance estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighting</td>
<td>Yes, good</td>
<td>Y, g</td>
<td>Y, g.</td>
<td>Y, but !(1)</td>
<td>Y, but !(1)</td>
</tr>
<tr>
<td>Prediction</td>
<td>Y, g</td>
<td>NO</td>
<td>Very limited</td>
<td>NO</td>
<td>Strongly misleading (2)</td>
</tr>
<tr>
<td>Proba law</td>
<td>Y, bad</td>
<td>Y, very bad</td>
<td>Y, quite good</td>
<td>Y, misleading</td>
<td>Strongly limited</td>
</tr>
</tbody>
</table>

(1) For partial non-response, however, each variable need its own set of weights. If the different variables are dedicated to separate univariate analysis, its make no particular problem except (but it is a practical important one!!) the fact to construct and store a great numbers of set of weights.

For (say) bivariate analysis, it becomes necessary to construct a specific set of weight (and to use the information contained in the ‘ears’ of the data). These weight, applied to the marginal distributions, have no coherence with the univariate weights (eg the marginal distribution in a two way table will not be the same than the distribution based on a univariate analysis). In practice, this coherency is a necessity which is achieved mechanically by imputation (whatever the method to fill the missing information: the goal is ONLY to get a rectangular ‘data’ set!).
(2) However, a rectangular data set with imputed data give a false appearance of good information. You cannot estimate the variance as if the imputed values were true values, especially when a prediction method is used.

Example: $s$ is a SRS of size 6: $(3,1,6,3,2,3)$.

Mean: $\bar{x} = 3$; ‘estimated’ variance of the mean

$$\frac{1}{6} \sum \frac{(x_k - \bar{x})^2}{5} = \frac{14}{30} \approx 0.467$$

Now we learn that data #1,4,6 are imputed by the mean. True variance estimation is $14/(3*2)=2.133$, inflated by a factor 5!

(3) The same is true when a proba law has been used, a bit attenuated (exercise: same data as above but hot-deck imputation). But there are two more crucial problems. First when we impute $\tilde{y}_k = \hat{y}_k + \varepsilon_k$, the random character of the $\varepsilon_k$ is not due to the sampling mechanism although the variance it induces is a true error! This extra variance can be very great (Exercise: Prove that hot deck imputation on a SRS with 1/3 (33%) hot-deck imputed data by SRS of the respondents increase the variance of the mean by 1/8=12.5%). We will see (with a little bit of luck!) in the last lecture some methods for limiting this waste...

Secondly, estimating a probability law is a much more risky task than finding a (scalar) predictor (the predictor may be robust against a lot of proba laws).
Weighting for non-response and generalized calibration

1-Model
The response mechanism is assimilated to a random sampling model conditional to \( s \). Let \( r_k \) be the (0,1)-variable indicating the response (conditional on \( s \)). The response probabilities \( P_k = E, r_k \) and moreover their inverse (non-response weights) \( g_k \) are the essential parameters to be estimated.

\[
This\ is\ typically\ the\ two\ phases\ sampling:\ U \rightarrow s \rightarrow r. \quad p(s) \rightarrow q(r/s)
\]

Technically, if you know perfectly how works the response mechanism, you are working with a two phases survey (or three if your initial survey was designed in two phases!): the first phase was perfectly controlled and the second is only observed and the parameters of the sampling mechanism are unknown.

Observe that the case of an ‘exhaustive’ survey is perfectly relevant (or the non-response correction in some take-all strata).

If we assume (which can be discussed, according to the data collection process) that response of the different units are independent, we get the Poisson sampling scheme as a model for the non-response mechanism.
Therefore:
$$q(r \mid s) = \prod_{k \in r} P_k \prod_{k \in s-r} (1 - P_k)$$

However some counts (number of respondents, by groups, ..) are observed and it is natural to condition under the observed data. In particular, variance estimation has to be conditional because it catches the actual precision of our estimates. **Ex:** Bernoulli → SRS

**2-Examples:** Parametric and non-parametric models

The set of $P_k$ ‘s is too big for a simple ‘direct’ estimation. From another point of view, generally, the response mechanism is never completely specified. One adopts a response model of the following form: $P_k^{-1} = g_k(z_k, \{z_{\ell \neq k}\}, s) = g_s(z_k)$ where $z_k$ is a set of explanatory variables which can contain some variables of interest observed on $r$ uniquely. If the $z$ are themselves too complicated to estimate non-parametrically the $g$ we use some parametrical model. We limit ourselves to this case in the following. This is complicated enough!

**Examples**: Generalized linear model.

- **Log – Linear**: $\exp(-z_k^T \beta)$
- **Log – Linearbis**: $1 - \exp(-z_k^T \beta)$
- **Logit**: $\exp(-z_k^T \beta)/(1 + \exp(-z_k^T \beta))$

...
Particular case 1: $z_k$ is the coding (0, ..., 1, 0) of a qualitative variable: $P(z_k' \beta) = P(\beta_i) = b_i$.
This leads to the homogeneous response-groups model.

Particular case 2: $z_k' = (I_k, J_k) = (0, ..., 1, ..., 0, ..., 1, ...)$ with a 1 in $i$ and $I+j$ th position.
The parameter is: $\beta' = (\alpha_1, ..., \alpha_i, ..., \beta_1, ..., \beta_j, ..., )$ \quad $\alpha_i = \log A_i$ \quad $\beta_j = \log B_j$.
If $c=i*j$ then $P_c=A_iB_j$. This is the formal raking-model.

These two last methods are very often used in practice.

**3- Estimation of the parameters**

3-1 Incidence of the use of estimated parameters for estimation

If the model is true (good variables, some of them could be variable of interest), then the estimation process leads to $\hat{g}_k = g_k + \delta_k$, estimation error. The natural estimator for a total is: $\hat{Y} = \sum_r \hat{g}_k d_k y_k$ (the $d$ are the initial weights on $s$, HT or calibrated). It can be easily shown that it is nearly unbiased ($O_p(1/n)$). The same kind of argument show that its (asymptotic!) variance is the same (up to small order terms) than the variance of the two phases sampling.

**NB:** In the case of a non-parametric estimation of the response probas, the conclusion is the same (although technically a bit more complex).
3-2 How to estimate the parameters of the response mechanism?

If we want to estimate the $\beta$ parameters in a classical way, we have to use explanatory variables present in the initial sample $s$ for the 'dependent variable' $1\{k \in r | k \in s\}$. This excludes variables present only in $r$ (and causing eventually the non-response!!).

This is the case for maximum likelihood (moreover the iid assumption is almost never justified!), least squares, chi-2 min, more or less generalized moments ($\sum_r z_k = \sum_s P_k z_k :$ you need the $z$ on $s$ !!), etc.

**Use of the calibration principle!**

a) Don’t modify a ‘good’ estimator! $\sum_r d_k x_k g'_k = \sum_s d_k x_k$.

b) If $X$ is known in the population, try:

$$ X = \sum_r d_k g'_k (\hat{\beta}) x_k $$

$(\text{dim}(x)=\text{dim}(\beta))$.

It’s looks like calibration equations and, technically, it is solved on the same way.

c) In some cases it will be:

$$ \begin{pmatrix} X_0 \\ \hat{X}_1 \end{pmatrix} = \sum_r d_k \begin{pmatrix} x_{0k} \\ x_{1k} \end{pmatrix} g'_k (\hat{\beta}) $$

Remark: In all cases the estimation error is $O_p \left( \frac{1}{\sqrt{n}} \right)$. 
4-Variance and estimating equations

In the process of estimating $\beta$ applying the calibration principle, you have a fantastic mathematical trick: put $\hat{\beta} = \beta + \lambda$ (say!):

$$
\begin{bmatrix}
X_0 \\
\hat{X}_1
\end{bmatrix}
= \sum_r d_k \begin{bmatrix}
x_{0k} \\
x_{1k}
\end{bmatrix}
g_k(\hat{\beta})
= \sum_r d_k g_k(\beta) \begin{bmatrix}
x_{0k} \\
x_{1k}
\end{bmatrix}
\frac{g_k(\beta + \lambda)}{g_k(\beta)}
= \sum_r d_k g_k(\beta) \begin{bmatrix}
x_{0k} \\
x_{1k}
\end{bmatrix} G_k(\lambda)
$$

The $G_k$ are true ‘calibration functions’ ($= \frac{g_k(\beta + \lambda)}{g_k(\beta)}$) and the equation just above is exactly a generalized calibration equation. If the model is true the $g_k(\beta)$ are the true response proba and the $g_k(\hat{\beta})$ are interpreted as the estimation of those probabilities; $\lambda$, the calibration parameter, is now interpreted as an estimation error on $\beta$ and is not identifiable.

We are now exactly in the context of generalized calibration, we have at hand all the elements for variance approximation and estimation. In particular the instruments for the global regression are $z_k^* = \frac{\partial G_k}{\partial \lambda} \bigg|_{\lambda=0} = \frac{g_k'(\beta)}{g_k(\beta)} = (GLIM Case) \frac{g'(z_k'\beta)}{g(z_k'\beta)} z_k$ (which have to be replaced by the estimation based on $\hat{\beta}$ in the numerical computations). ($z_k^* = z_k$ only in the exponential case!).
The approximate variance is given by $Q_1(e^U) + Q_2(e_U)$ where $Q_1$ is the full response variance, $Q_2$ the non-response variance, $e = y - \tilde{B}_0 x_0 - \tilde{B}_1 x_1 = \text{Residuals}(y/(x_0, x_1) \text{ using } z^*)$ and $e^U = y - \tilde{B}_0 x_0$. In the case of a Poisson response model the variance is reduced because the parameter estimation process leads to some kind of conditional inference. For instance if the response model is Bernoulli (Poisson+equal proba), the approximate variance becomes as using SRS having for size the number of respondent.

Variance estimation of the second part ($Q_2(e_U)$) is generally straightforward. Estimating the first part ($Q_1(e^U)$) is complicated because we have to start from the respondents and to estimate an estimator of $Q_1(e^U)$! However if the response model is Poisson there are interesting simplification.

In this case the ‘reverse approach’ postulating that the response mechanism is independent of the sampling process makes the things quite simple.
5-Examples

-Example 1 (exercise): Poissonnian Response Mechanism: $x_k > 0 \ P_k = x_k \beta$ (probability proportional to size!). The calibration principle leads to (if $N$ is not known):

$$\sum_r \frac{x_k}{\pi_k x_k \beta} = \sum_s \frac{x_k}{\pi_k} \Rightarrow \hat{\beta} = \frac{R^*}{\hat{X}} \text{ with } \left( R^* = \sum_r \frac{1}{\pi_k} \right)$$

According to the fact that we have a ratio estimator (conditional to $s$) the (approximate) 2\textsuperscript{nd}-phase variance is equal to:

$$E \sum_p \sum_s P_k \left(1 - P_k\right) \left(\frac{y_k - x_k \frac{Y}{X}}{\pi_k P_k}\right)^2 \text{ estimated by } \sum_r \left(1 - P_k\right) \left(\frac{y_k - x_k \hat{Y}_{exp} / \hat{X}_{exp}}{\pi_k P_k}\right)^2$$

Remark: this is the exact variance in the case of an exhaustive survey.

-Example 2: The Homogenous Response Cell model and formal poststratification. $g_k(\beta) = \beta_h$ if $k$ belongs to the cell $U_h$. The ‘natural’ estimator’ is $m_h/n_h$. Using the calibration principle, if $N_h$ is unknown, we get $\hat{N}_h / \sum_{rh} 1/\pi_k$ and if they are known, the final estimator is $\sum_{h} N_h \hat{Y}_h$, the USUAL practice.

Exercise: Estimator and variance estimation if some $N_h$ are known and some unknown?
-Example(s) 3: GLIMs

- **Log-linear model**: \( g_k = \exp(x_k'\beta) \): The ‘calibration estimating equations’ are exactly the calibrating equations of the raking-ratio of \( r \) on \( s \); that is the calibration in the second phase of the sampling. This gives a justification of the ‘Formal Raking-Ratio’.

- **Log-linear bis**: score = calibration!

- **Logit**: score = moment. How to estimate a logit model using calibration?

\[
\sum_r d_k x_k = \sum_r d_k x_k \frac{1 + \exp(z_k'\beta)}{\exp(z_k'\beta)} \quad \Rightarrow \quad \sum_r d_k x_k \exp(z_k'\beta) = \sum_r d_k x_k
\]

You have get calibration equations for the exponential/Raking Ratio method a bit special because of the second member of the equations. The last equation makes clear the importance of having the constant among the \( z \). It can be rewritten

\[
\sum_r d_k x_k \exp(-\beta_0 - z_k'\beta_1) = \sum_r d_k x_k \quad \text{or} \quad \sum_r d_k x_k \exp(-z_k'\beta_1) = \exp(\beta_0) \sum_r d_k x_k
\]

If 1 is also in the \( x \), \( \beta_0 \) (size) and \( \beta_1 \) (structure) are estimated independently.
-Example 4: The \( x \) vector correspond to a classification of the units available in the frame, for instance the situation last year of the same classification to day, encoded by vector \( z \), available only for the respondents (in the practice, in French surveys, social status of the household, which is a known factor of non-response).

\[
x'_k = (0,0,0,0,1,0,...,0); \quad N_i = [N_i] \quad i \in I; \quad z'_k = (0,0,0,0,1,0,...,0); \quad Y_a = \sum_{r,a} y_{k,r,a} \quad \text{vector} Y_A; \quad \beta = (..., \beta_a,...).
\]

(frame situation) \hspace{1cm} (present situation)

Let \( R_{ia} = \sum_{r} x_{k} z'_{k} = \text{number of units classified in } i \text{ in the frame and } a \text{ in the survey.} \)

Estimating equations are: \( N_i = \sum_{a} R_{ia} (1 + \beta_a) \)

Variance: \( \sum_{a} \beta_a (1 + \beta_a) \sum_{i} \sum_{k \in i} (y_{k} - \hat{Y}_i)^2 \)

Remarks: For simplicity we have treated the case of an exhaustive survey. Adaptation is quite easy for the general case.

Observe that the \( z \) variable is not observed for non-respondent! In fact, the total of this variable (number of units classified in the \( a \) cell) can be a statistics of interest. We are typically in a case of ‘non-ignorable’ non-response. In practice, the method works well. (The worker representing person case).
Example 5: worker delegates, drug consumers, people voting for FN,...binary variates:

Data:

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Non-response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>.6</td>
<td>.1</td>
<td>.3</td>
</tr>
<tr>
<td>Blue</td>
<td>.2</td>
<td>.3</td>
<td>.5</td>
</tr>
</tbody>
</table>
6-Some remarks about a ‘good’ response model

What has been said for the logit model can be slightly generalized according to the logic of the correction. Observe first that we can always include the constant in the instruments and that a constant non-response rate should always be admissible. This leads to a specification of $g$ having the form

$$g_k(\beta_1, \beta_+) = \beta_1 g_{+k}(\beta_+) \text{ with } g_{+k}(0) = 1.$$  

Moreover the constant can also be included in the calibration variables because at least $\sum_s d_k x_k$ is known. This is sufficient to estimate the scale factor, the other variables tackling the deformation of structure. It is seen simply by the fact that the first calibration equation is written as:

$$N(\text{or } \bar{N}) = \sum_r d_k \beta_1 g_{+k}(\beta_+)$$

and the other as:

$$\sum_r d_k g_{+k}(\beta_+)(x_k - \bar{X}).$$

The structure parameters are estimated independently of the size parameter as for the logit model.
Deterministic imputation, prediction estimation, and calibration

1-The estimator and the superpopulation model

Imputation = replace the missing values in \( o=s-r \) by a ‘plausible’ one \( \hat{y} \).

Estimator for a total \( \hat{Y}_{imp} = \sum_s w_k \hat{y}_k = \sum_r w_k y_k + \sum_o w_k \hat{y}_k \).

The same for functions of totals and non-linear quantities.

The plausible value is a prediction \( \hat{y}_k \) of \( y_k \). We have to use a superpopulation model \( m \) such that \( E_m y_k = \hat{y}_k \) to derive a predictor. Using the ‘soft model’, the \( (y_k, x_k; k \text{ in } U) \) are independent realisations of some probability law on \( \mathbb{R}^{p+1} \). The \( x \) are observed on \( s \) and \( y \) on \( r \) uniquely. Making the model a bit less ‘soft’, we assume that the conditional laws (given \( x \) ) exist, are regular and vary smoothly. The function \( f(x) = E_m(y \mid x) \) is estimated non-parametrically (if \( r \) large enough and the size of \( x \) small enough) or parametrically: \( E_m(y \mid x) = f(x, \theta) \), \( \theta \in \mathbb{R}^p \).

Examples:

- **Cold deck**: \( \hat{y}_k \) is some given variable (in the frame, situation last year, expert value…).
- **Imputation by rule**: It is a variation of the cold deck method. Ex: if age<16 => activity=no;
- **By the mean**: impute the mean value observed among the respondents. The model is the \( y_k \) are i.i.d.
**Ratio**: Some variable $x$ (numerical, positive) is observed on $s$. The model is $y_k = Rx_k + \varepsilon_k$.

**Regression**: Linear model on $r$ using covariates observed in $s$ and use the prediction formula.

**Logit or generalized linear model**: Used for the ‘expectation’ of a $(0,1)$ variable.

**By class**: The last four method are very often applied by class. In fact, it is often useful/necessary to use different methods in different classes.

**Non parametric**: Use a ‘smoother’ (kernel, local polynomial ) adjusting a function $\hat{y}_k = f(x_k)$.

*In the following we focus on the parametric case.*

**Remark**: It is more or less assumed, for most methods, that the variables are real (numerical). In practice, one generally use different methods for numerical, $(0,1)$, ordered (or small numerical: number of children), or categorical variables. For ordered variables, it is possible to use prediction methods but with the drawback to impute non admissible values. In particular, for $(0,1)$ variables, it is sometimes convenient to impute a number between 0 and 1 interpreted as the proba to get a 1. More often, people prefer a random imputation in this case, with all the difficulties bound to this technique (see later).
2-Derivation from a superpopulation model

In a very general way, the $\theta$ parameter is defined at the population (or $U$) level by a max (or min!) additive criterion $\theta = \text{argmin}(\sum_U L_k(\theta))$, the $L_k$ being a numerical function associated to the individual $k$ and having for arguments at least $y_k$ and $x_k$.

We now have to distinguish four(!) levels in the problem of estimating. The central one is the population (or $U$) level. We have then the $s$ level where the plug-in estimator verify:

$$\forall s \theta_s = \text{argmin}(\sum_s d_k L_k(\theta))$$

Then there is the $r$ level where the plug-in estimator verify:

$$\forall r \theta_r = \text{argmin}(\sum_r g_k d_k L_k(\theta))$$

At this level we have the information allowing to estimate $\theta_r$ and therefore $\theta$.

**Remark**: We need to have a model for the response mechanism for a correct answer. Often, in practice, SRS, or SRS inside cells, is implicitly assumed. Imputation is NOT a method allowing the economy of neglecting the response mechanism!

Eventually there is also the $m$ (for model) level where the parameter is defined by:

$$\theta_m = \text{argmin}(\int_{\mathbb{R}^{p+1}} L_k(\theta) dM(k))$$

where $M$ is the measure supporting the superpopulation model.
3-Estimating equations for parameters and variance

We assume that \( \hat{\theta} \) is solution of the estimating equations (score type):

\[
\forall r \quad \sum_k d_k g_k \ell_k (\hat{\theta}) = 0 \quad \text{where} \quad \ell_k (\theta) = \frac{\partial L_k}{\partial \theta} \in \mathbb{R}^p.
\]

Most of the time this set of \( p \) equations take the form of normal equations (regression terminology) with:

\[
\ell_k (\theta) = z_k (y_k - f_k (\theta))
\]

or, as for regression or GLM like logit where (for the moment) the \( z \) are arbitrary \( p \)-dimensional instruments:

\[
\forall r \quad \sum_k d_k g_k (\hat{\beta}) z_k (y_k - f (x_k, \theta)) = 0
\]

This means that the variance of this expression is equal to 0, and we get the variance of \( \hat{\theta} \) by linearization:

\[
Var(\hat{\theta}) = Var(\sum_U \frac{\partial \ell_k}{\partial \theta}^{-1} \sum_r d_k g_k \ell_k (\theta))
\]

because \( \hat{\theta} \) is the substitution (plug-in) estimator of \( \theta \) in the equation \( \sum_U \ell_k (\theta) = 0 \) (with straightforward adaptations if working conditional to \( s \)).

If the model is true, the estimator is (asymptotically) unbiased for the total of \( y \) or of semi-linear quantities like \( \sum_U c_k y_k \) where the \( c_k \) are fixed constants like \( 1(k \text{ in a domain } D) \).
Up to second order terms we can write: 

\[ \hat{Y}_{imp} = \sum_s d_k f(x_k, \theta) + \sum_s d_k f_\theta'(x_k, \theta) (\hat{\theta}_r - \theta). \]

As, conditional to \( s \), \( \hat{\theta}_r \) is nearly unbiased, we have 
\[ E(\hat{Y}_{imp}|s) = \sum_s d_k f(x_k, \theta). \]

We obtain now easily the variance:

\[ Var(\hat{Y}_{imp}) = Var(\sum_s d_k f(x_k, \theta)) + E(Var(\sum_s d_k f_\theta'(x_k, \theta) (\hat{\theta}_r - \theta)|s)). \]

The first term is not \( Var(\sum_s d_k y_k) \) (happily for making feasible variance estimation, don’t you mind?) and has a smallest variance! It can be estimated by replacing \( \theta \) by \( \hat{\theta}_r \).

The second term is estimated using the linearized variable for \( \theta \) AND the model of sampling for non-response.

**4-Choice of the instruments**

**4-1Standard considerations**

- Instruments must have the same dimension as \( \theta \).
- One of the coordinates (explicitly or implicitly) must be \( I \) such that the part of the estimator coming from \( r \) does not change: 
\[ \sum_r d_k y_k = \sum_r d_k f(x_k, \hat{\theta}). \]
- The $x_k$ (eventually weighted) are natural candidates, as it comes to be in ‘likelihood like’ derivation of normal equations.

4-2 Using some trick avoiding to compute and to use explicitly the response weights. Interesting possibilities consist in using the explanatory variables of the response model. Suppose, first, that we work with a with $q$-cells response model and let $z$ be the vector of their indicatives. The above equations

$$
\sum_{r} d_k g_k(\hat{\beta}) z_k (y_k - f(x_k, \theta)) = 0
$$

are clearly equivalent to

$$
\sum_{r} d_k z_k (y_k - f(x_k, \theta)) = 0
$$

and there was no need to compute the response probabilities.

If $q<p$ you can always complete the number of instruments (among the $x$ for instance). However if $q>p$ you cannot eliminate the bias completely with this trick.

More generally, suppose that our response model is a ‘semi linear generalized’ of the type described in Weighting-6 $g_k(\beta_1, \beta_+) = \beta_1 g_+(z'_k \beta_+) \text{ with } g_+(0) = 1$ (it seems not to be completely mandatory!). Then

$$
g_k \approx \beta_1 (1 + g'_+(0) z'_k \beta_+) = (\beta_1, \beta'_+) \left( \begin{array}{c} 1 \\ z_k \end{array} \right).
$$

The estimating equations

$$
\sum_{r} d_k z_k (y_k - f(x_k, \theta)) = 0
$$

are nearly unbiased because $g_k$ is a linear combination of $z$.

However a complete computation of the $g$ seems to remain necessary for variance estimation.
5-Examples: ratio imputation, regression imputation

Ratio imputation: \[ \hat{\theta}_r = \frac{\sum_{r} d_k y_k}{\sum_{r} d_k x_k} \quad \Rightarrow \text{Var}(\hat{Y}_{\text{imp-ratio}}) = \theta^2 \text{Var}(\sum_{s} d_k x_k) + E(\hat{X}_s^2 \text{Var}(\sum_{r} y_k - \theta x_k | s)). \]

Variance estimation follows naturally: \[ \hat{\theta}_r^2 Q_s (x_s) + \hat{X}_s^2 Q_r ((y - \hat{\theta}_r x)_r). \]

Regression imputation: Exactly the same kind of calculation leads to:
\[ \text{Var}(\hat{Y}_{\text{imp-regression}}) = B'\text{Var}(\sum_{s} d_k x_k)B + E(\hat{X}_s' \text{Var}(\hat{B}_r - \hat{B}_s | s)\hat{X}_s). \]

Variance estimation follows easily:
\[ \text{Var}(\hat{Y}_{\text{imp-regression}}) = \hat{B}'\text{Var}(\sum_{s} d_k x_k)\hat{B} + \hat{X}_s' \text{Var}(\hat{B}_r - \hat{B}_s | s)\hat{X}_s. \]
6-Summary:
If the imputation model is ‘true’, the imputed estimator by prediction is nearly unbiased for the total of $y$ and for ‘semi-linear’ quantities like domain. It remains also true for the estimation of mean on domain for instance. When a calibration equation is used for estimating the parameter, variance is reduced and can be estimated quite easily.

If no use is made of a response model, uniform response probabilities are implicitly assumed. Generally a bias is generated except in the case where the prediction parameter $\theta$ is estimated using normal equations and instrumental variables containing the variables causing the non-response.

However, when we are interested in variables derived from $y$ non-linearly, like $1(y_k < a)$, the estimator becomes inconsistent even for the total. In fact, the distribution of the $y_k$ in the imputed sample is strongly perturbed and false! The typical example is given by the ‘mean imputation’, which is the simple case of prediction imputation. It remains true in the general case.
Random imputation and balanced sampling

0-Imputation ‘Weighting Like’

Just a gadget! For simplification (but it happens that it was a real case!) suppose we have an exhaustive survey with some complete non-response. The best solution is to weight and we have solved our calibrations/estimations equations: 

\[ X = \sum_{U} x_k = \sum_{r} F_k(\hat{\beta})x_k = \sum_{r} w_k x_k. \]

Among the \( x \) variables, there is the ‘free of charge’ variable equal to 1, such that the sum of the \( w_k \) is equal to \( N \), the size of the population. Observe also that the \( w \) are all greater than 1. It happens that you cannot use the weights (in the practice, because the firm where the student had his training period used some exotic software allowing only simple sums…), but only imputation. You can do the following; we have \( w_k = n_k + \pi_k \) with \( n_k \) integer and \( 0 = <\pi_k < 1 \). Generally you have \( n_k = 1 \); if it is not the case, duplicate unit \( k \ n_k - 1 \) times.

Then, duplicate the \( k \) in \( r \) with probability \( \pi_k \) and verifying the constraints: (Cube!)

\[ X - \sum_{r} n_k x_k = \sum_{r} \pi_k x_k. \]

In particular, the sample will have the good size!
Moreover (Exercise, the last!), the variance of $\sum_{U} \tilde{y}_k$, the imputed estimator, is quite exactly the variance of $\sum_{r} w_k y_k$, the weighted one!

Rem: A very similar idea has been applied in the British ‘One count census’ (R.Chambers.)

1-Generalities

1-1 Some new problems
Prediction imputation produces a false distribution of the imputed variable. Under the same regularity hypothesis, a natural idea could consist in choosing a value at random in the conditional law given $x$. It requires a more precise (and more risky!) specification of the link between $x$ and $y$ in the superpopulation model. The kind of model to use differs according to the type of variable. It can be parametric (some law coming from the supermarket!), non parametric (the empirical law of some set of respondent for instance), or often semi-parametric (the predictor is parametrically estimated but the law of the residuals is non parametric).
Basically, we are in a situation where a superpopulation model, the same as in DetImp 2 has been estimated for $y_k$ conditional to the vector of covariates $x_k$. The difference consists in the fact that we estimate the full law, and not only the expectation of $y$ conditional on $x$. This is in fact necessary when $y$ is a ‘nominal’ variable. In the case of a numerical variable we have to estimate the law of residuals $\varepsilon = y - \hat{y}$. If $y$ is real it is habitual. If $y$ is (0,1) valued, or integer it is a bit special. However, in all cases, the estimations are based on the set of respondents.

The major problem with random imputation is that the random process of imputation, created by the statistician (ourselves!), with its random numbers, add an artificial variance to the estimators. This extra-variance can be large as it can be seen with the following example/exercise, not at all unrealistic:

Exercise/example: A SRS of size $n$ has been drawn; there are only $m$ respondents. As $m>n/2$, we draw a simple random sample of respondent of size $n-m$ for imputing the missing value. Let $\bar{y}_r = \frac{1}{m} \sum_{r} y_k$ the respondent mean, $V$ its variance, $\bar{\bar{y}}_r = \frac{1}{n} \sum_{r} \bar{y}_k$ the mean of the imputed sample, $\bar{V}$ its variance. Show that $\frac{\bar{V}}{V}$ is maximum when the non-response rate is 1/3 and take the value 1.125=1+1/8. What happens if the response rate is 50%? Explanation?
1-2-Quick inventory of methods

a- For quantitative variables:
\[ \tilde{y}_k = \hat{y}_k + \varepsilon_k \] with the \( \varepsilon \) coming from a ‘textbook’ law (Gaussian?) whose parameters have been estimated. Ex: \( \hat{y}_k \) comes from a regression (or a ratio, or a mean) and \( \varepsilon \) is Gaussian.

b- The same as a- but the residuals are non-parametric: for \( k \) in \( r \)
\[ \varepsilon^*_k = \frac{y_k - \hat{y}_k}{\sigma_k} \] for \( k \) given to \( l \) in \( o \), \( \varepsilon_l = \varepsilon^*_k \sigma_l \).

c- Donnor imputation:
Classical hot-deck: impute a random donor= law is the same for all units = impute the mean+ a centred random residual coming from the empirical law (nonparametric estimation of the law).
Hot-deck by class: the law is the same for all members of the class.

d- Nearest neighbour method.
Construct a distance on the covariate \( x \) and impute \( \tilde{y}_i = y_k \) for the \( k \) such that \( \text{distance}(x_k, x_l) \) is minimum. Despite its appearance, this methods must be seen as a random imputation.

e- for a qualitative variable , hot deck by class is widely used with the well-known ‘table algorithm. See also later ( part 3-2).
2-Formalization and Estimation Theory

2-1 The extra-variance

Formally, in the case of a numerical variable (including \((0,1)\) or integer), we can write 
\(\tilde{y}_k = \hat{y}_k + \varepsilon_k\), sum of a predictor, which would be used if we had chosen the prediction method, 
and of a random centred variable, independent of the sampling/response process. 
The predictor \(\hat{y}_k\) depends only on \(r\) and not of our randomisation technique. 
Therefore, for the estimation of the total of \(y\):

\[
E(\hat{Y}_{\text{imp-random}}|r) = \hat{Y}_{\text{imp-pred}} \quad \text{and} \quad \text{Var}(\hat{Y}_{\text{imp-random}}) = \text{Var}(\hat{Y}_{\text{imp-pred}}) + \mathbb{E}(\text{Var}(\hat{Y}_{\text{imp-random}}|r))
\]

So, the first term is the (quasi-randomisation) variance of the prediction method. The second one 
correspond to the extra variance (imputation variance), that is the variance of \(\sum o d_k \varepsilon_k\).

*What for is it the price to pay? Is it only aesthetic of a good looking distribution?*

In fact we have a deeper property. Suppose we have to estimate the total of some (non-linear) 
transformation of \(y\) say \(t_k = \varphi(y_k)\). For instance \(t_k = 1(\hat{y}_k < a)\) for the distribution function. If the model is true, that is if \(y_k\) follows exactly the conditional law described by the model (and if, 
moreover, our estimation of the law is sufficiently ‘good’), then \(\varphi(y_k)\) and \(\varphi(\hat{y}_k)\) have the same
law. Therefore $\sum_{s} d_k \varphi(y_k)$ and $\sum_{s} d_k \varphi(\hat{y}_k)$ have the same expectation under the imputation model.

In fact, if we refer to the discussion in *DetImp2* and the four level of ‘sampling’ (superpopulation, finite population, design sample, respondent sample) the superpopulation level makes appear clearly that we are indeed interested in the set of $y$-measurable functions and their distribution functions. In the case or real (eventually vector valued) variable this vector space is infinite dimensional. As we have to estimate such quantities starting from a finite set of values, we must use some kind of finite dimensional approximation for those distributions. In the case where $y$ generate a finite $\sigma$-field it seems that things could be simpler (see below).

The overall conclusion is that, under quite heavy hypothesis, random imputed estimators works also for non-linear transformation of $y$, in particular for the distribution function. The variance can be computed in the same way that for a total, if we are able to predict the expectation of $\varphi(y_k)$ under the imputation model.

### 2-2 Exercise (or example!)

Continued from 1-1: *Suppose you have to estimate the median or some other quantile. Show that the mean imputation is irrelevant. Show that hot deck imputation is nearly unbiased, with an extra variance compared with the weighting method. Try to find the most efficient imputation method (Hint: in the case of a non-response rate with value 1/3, use of small strata with two units).*
2-3 General recipes for the extra variance

The extra variance (imputation variance), that is the variance of \( \sum_o d_k \varepsilon_k = \sum_o d_k^2 \sigma_k^2 \) if the draws are independent. To reduce this variance, we have to introduce negative covariance between the \( \varepsilon_k \). A general way to do that is to introduce constraints on the imputed residuals. One family of methods consists in using a balanced sampling (cube or not cube). In fact, we will essentially give examples until the end of the course, the topics being too new for a general theory. One will observe that when the Cube (say) method is used instead of Poisson sampling, a part, sometime great, of the imputation variance is removed.
3-Applications, Examples and Concluding Remarks

3-1- Imputation of a (0-1) variable and balanced sampling

\( y_k = 0 \) or \( 1 \) and \( Q_k = Pr(y_k = 1) \) has been estimated consistently. The first idea is to consider the \( y_k \) as independent conditional to \( x_k \). This lead to impute the \( 1 \) values by a Poisson sampling.

*(Rem and exer : Observe that, by no mean, we will impute the nearest admissible value; What happens if \( Q_k = .9 \) for all units? What would you do if you had to bet on the result of imputation?)*

There is an obvious drawback: the best predictor for \( \Sigma_0 y_k \) is \( \Sigma_0 Q_k = \nu < n-m \); we can assume (by rounding ) that this number is an integer, and it is natural to impose as a constraint \( \Sigma_\theta \hat{y}_k = \nu \).

This mean that we have to perform a fixed size unequal probability sampling, a conditional Poisson for instance. As it was seen, \( Var(\Sigma_\theta \hat{y}_k) = \Sigma_\theta Q_k (1-Q_k) \) for Poisson , and of course \( 0 \) for fixed size sampling. So we have (in this case drastically!) reduced the imputation extra-variance.

It is possible to do better by imposing other constraints. For instance, suppose that the \( Q_k \) has been estimated parametrically by some estimating equation set having the form (more or less a calibration!) \( \Sigma_r z_k (y_k - Q_k(\theta)) = 0 \). If you want to keep unchanged this estimation for the sample \( s \), and if the \( z_k \) are also in \( s \), you have to satisfy the constraints \( \Sigma_{o=s} z_k (y_k - Q_k(\theta)) = 0 \), that is to perform a balanced sampling. The imputation variance will be reduced. For the total of \( y \), it is reduced to \( 0 \) if the weights \( d_k \) are included in the \( z_k \).
Remark: The constraints can be arbitrarily chosen (not necessarily bound to the estimation process). Only information on $s$ is required.

Now, we can also be more precise on variance and variance estimation: the imputation variance in estimating the total of some $ω_k$ is the variance of the residuals of the $d_kω_k$ on the $z_k$ using the weights $\frac{(1-Q_k)}{Q_k^2}$, according to the theory for the variance of a balanced sampling. Observe that $z_k$ can be chosen arbitrarily; this fact opens many perspectives….

3-2 Qualitative variable and balanced sampling of cells
For each $k$ in $o$ we have to impute some modality $i$ of a qualitative variable having $I$ modalities. We assume that the probas $Q_{ki}=Pr(y_k=i)$ have been estimated with some reliable model, for instance using a log-linear adjustment $Q_{ki}=exp(x_k'q_i - c_k)$ on the respondents. Assume, moreover, that $\sum_{k \in o} Q_{ki} = ν_i$ is a vector of integer (this can be done by ‘raking’ conveniently the probas). Our problem is to fill the table $oxI$ with $y_{ki}$ being zeros or ones verifying the two sets of constraints:

$$\sum_{k \in o} y_{ki} = ν_i \text{ and } \sum_{i \in I} y_{ki} = 1.$$ 

In fact, it is a problem of balanced sampling (of cells!). The cube method can be utilized.
**Remark:** An heuristic method can be the following: impute the first unit with the given probas; Adjust by raking the remaining units remarking that $\nu_i$ does not change for the columns where a zero has been attributed, and $\nu_i$ becomes $\nu_i - I$ for the ‘winning’ cell.

However we can do better! Like in section 3-1, we can add other constraints (related to the estimating equations of the $P_{ki}$ for instance). The imputations has to meet conditions looking like $\sum_{k \in o} z_k \tilde{y}_k = \sum_{k \in o} z_k Q_k$ denoting as row-vectors the ‘vectorialized’ $\tilde{y}_k$ and $Q_{ki}$ (the preceding quality is in fact between matrixes). Here also the Cube method can be used to perform a balanced imputation, with the benefits already described in section 3-1: invariance of the estimation of the imputation model, reduction of the imputation variance for some estimators.
REMARK: Is random imputation really mandatory in 3-1 and 3-2?
From a naïve point of view it seems that only admissible value could be imputed. In fact it is not really true if we have in mind what we want to do. Referring to 2-1, what we want to do is to be able to deal with an arbitrary function \( \varphi(y) = \varphi(i) \) with \( i \in I \), a finite set. Let \( \varphi \) be also the vector of the \( \varphi(i) \) and \( Q \) the vector of the (estimated) probabilities. The imputation prediction total of \( \varphi(y) \) is estimated by:

\[
\sum w_k \varphi(y_k) + \sum w_k \varphi' \hat{Q}_k = \sum w_k \varphi' \hat{Q}_k \quad \text{if we have used convenient estimating equations for } \hat{Q}_k .
\]

All properties described in \textbf{DetImp3-4} are valid without the extra parasitic variance caused by random imputation. The only practical bound is the size of \( I \).
3-3-Quantitative variable and balanced sampling of residuals

We have to impute quantities like $\hat{y}_k = \hat{\gamma}_k + \text{random}$. The random terms take the form $\sigma_k e_k^*$ where the $e_k^*$ follows the estimated probability law of the imputation model. The $\sigma_k$ are variables known on the whole sample $s$. This can take many patterns, from the banal normal assumption to an estimated nonparametric law. Anyway imputing independent $e_k^*$ in the imputations provides always the extra parasitic variance $\sum_o d_k^2 \sigma_k^2$. The idea is to reduce the imputation variance by imposing a zero correlation between some set of variables $z_k$ and the residuals: $\sum_o z_i e_i^* = 0$.

Three techniques can be used according to the model.

1-Gaussian model.
The residual follows a $\text{Normal}(0,1)$ law. It is always possible to find a multivariate Normal for $\varepsilon_k$, $k \in o$ with $\text{Normal}(0,1)$ marginal’s and verifying $\sum_o z_i \varepsilon_i = 0$ (not easy result but true!).

The two other techniques are based on a non-parametrical estimation of the law of residuals: collect the normalized empirical prediction errors in the estimation procedure; that is, for $k$ in $r$ $e_k^* = \varepsilon_k / \sigma_k$ ($k \in r$), with weights proportional to $d_k g_k$ adding to $l$. 
2- For me, I like the following technique (for details try to find the acts of the JSM meeting 2006-Seattle...and good luck!).

It is a kind of balanced sampling (different of the Cube) in the population of empirical residuals in $r$. The algorithm is something like that:

- $o$ is sequentially ordered $l=1$ to $n-m$.
- Choose the first $\varepsilon^*_k$ at random.
- Then, for $l>1$, if $z'_l(\sum_{i=1}^{l-1} z_i \varepsilon^*_i) > 0$ draw $\varepsilon^*_i$ at random in the negative part of the distribution, in the positive part in the opposite case.

In can be shown the partial sums remains $O_p(1)$ (instead of $O_p(l^{1/2})$) all along the process, that means that we achieve balance. In particular, if one of the coordinates of $z$ is $d_k \sigma_k$, the estimation of the total of $y$ remain nearly the same. More generally, if the predictors have been adjusted using normal type equations, the fact to reuse the same instruments insure that the estimating equations for the parameter will have the same solution with the imputed data as with the respondent data. In particular a regression estimator adjusted on $r$ will not be altered by the imputation process.
Here is an illustration with nice simulations. The upper curve describes the evolution of the norm of the vector $S = \sum_{i=1}^{\ell-1} z_i \varepsilon_i$ when the imputations are independent, the other one is with control. The same sequence of random numbers is used in the two cases. This is to see on the figures: large jumps are at the same places on the two curves.
3- A way to tackle the problem is to use the Cube (joint work with G.Chauvet and D.Haziza).
The goal is to select cells in the table with inclusion probabilities $p_j (j \in r)$. The constraints are:

- one cell in each row and total of the $\sigma_k \epsilon_j (k \in o)$ equal to 0.

- A matrix of constraints:

<table>
<thead>
<tr>
<th></th>
<th>j=1</th>
<th>j=2</th>
<th>j=3</th>
<th>j=4</th>
<th>j=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=2</td>
<td></td>
<td>$\sigma_k \epsilon_j \ p_j$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here is the matrix of constraints:

<table>
<thead>
<tr>
<th></th>
<th>1-1</th>
<th>1-2</th>
<th>1-3</th>
<th>1-4</th>
<th>1-5</th>
<th>2-1</th>
<th>2-2</th>
<th>2-3</th>
<th>2-4</th>
<th>2-5</th>
<th>3-1</th>
<th>3-2</th>
<th>3-3</th>
<th>3-4</th>
<th>3-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k=2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>k=3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Varcal</td>
<td>$\sigma_1 \epsilon_1$</td>
<td>$\sigma_1 \epsilon_2$</td>
<td>$\sigma_1 \epsilon_5$</td>
<td>$\sigma_2 \epsilon_1$</td>
<td>$\sigma_2 \epsilon_2$</td>
<td>$\sigma_2 \epsilon_5$</td>
<td>$\sigma_3 \epsilon_1$</td>
<td>$\sigma_3 \epsilon_5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$&lt;$</td>
<td>$/$ $p_j \ j = 1 à 5$</td>
<td>$&gt;$</td>
<td>$/$ $p_j \ j = 1 à 5$</td>
<td>$&gt;$</td>
<td>$/$ $p_j \ j = 1 à 5$</td>
<td>$&gt;$</td>
<td>$/$ $p_j \ j = 1 à 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Then use the cube method (a bit modified for a quick computation). At the end of the flight phase, we get only 4 non-zero probas and a rounding involving only one individual.
REMARK: Is random imputation really mandatory?

Unhappily, generally YES.
The same argument as after 3-2 could be used to say NO. If we are interested in the total of $\varphi(y)$ we can say that $\hat{y}_k = \varphi(\hat{y}_k + \sigma_k \varepsilon_j)$ with probability $p_j$ and we are exactly in the case of the preceding remark. This miracle is due to the approximation of the infinite by the number of respondents! However we have to create for each $k$ in $o$ a vector having for size the number of respondents, which seems a bit too much...
3-4-Summary and final remarks

SI LE MODELE EST VRAI, l’imputation avec aléa fournit des estimateurs consistants et approximativement sans biais de fonctionnelles non linéaire de $y$. Elle produit une variance supplémentaire qu’on peut réduire plus ou moins en utilisant des techniques d’échantillonnage avancées. L’utilisation de données imputées est cependant toujours risquée, spécialement s’il y a des aléas, car on utilise alors une loi estimée et non pas un simple prédicteur. En particulier l’imputation respecte très mal les corrélations entre variables de sorte que l’utilisation de données imputées est presque impossible pour la modélisation économétrique.

Ne jamais oublier qu’on n’a créé aucune information. Parfois, au contraire, on en a détruit (dépendances entre variables dans l’imputation par prédiction).

Il ne faut jamais estimer naïvement la variance d’une estimation comportant des valeurs imputées (si VOUS n’êtes pas naïf, malgré tout, vous pouvez utiliser un estimateur naïf comme point de départ, car tous les logiciels connaissent bien les estimateurs naïfs !).

Post–scriptum : Please, DO NOT SPEAK TO ME OF MULTIPLE IMPUTATIONS!!