La coordination d'échantillons d'enquêtes auprès des entreprises





Fabien Guggemos, Olivier Sautory
INSEE

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Numéros aléatoires

Technique des numéros aléatoires utilisées par de nombreux INS pour coordonner des échantillons auprès des enquêtes entreprises.

A chaque unité k de la population U (y compris aux nouvelles unités) est attribué, indépendamment des autres, un nombre ω_k, tiré aléatoirement selon une loi uniforme sur [0,1 [.

Tirage de Poisson: sélection des unités k dont les numéros aléatoires sont dans l'intervalle [d, d + π_k[, π_k = probabilité d'inclusion de l'unité k.

Tirage aléatoire simple (TAS) : sélection des n unités ayant les les plus petits numéros aléatoires ω_k supérieurs à d.

Définition d'une fonction de coordination

Fonction de coordination g = fonction borélienne (i.e. mesurable) de [0,1[vers [0,1[, déterministe, qui préserve la probabilité uniforme :

Si P est la probabilité uniforme sur [0,1[, alors la probabilité image doit vérifier:

$$P^g = P.$$

→ Notamment, pour tout intervalle I = [a,b [inclus dans [0,1[:

$$P(g^{-1}(I)) = P^g(I) = P(I) = b - a$$

→ Et en particulier…
De même que les $(\omega_k)_{k\in U}$, les numéros aléatoires transformés $(g(\omega_k))_{k\in U}$ sont indépendants et suivent une loi uniforme sur [0,1[

Sélection des unités

Pour ***chaque*** unité k dans le champ d'intérêt :

- un numéro aléatoire ω_k dit permanent, (invariant dans le temps)
- une fonction de coordination $g_{k,t}$ qui évolue à chaque nouveau tirage d'échantillon t = 1, 2, …

1. Tirage de Poisson

Sélection des unités k telles que $g_k(\omega_k) \in [0, \pi_k[$
où π_k = probabilité d'inclusion de l'unité k

$$P(k \in S) = P(g_k(\omega_k) \in [0, \pi_k[) = P^{g_k}([0, \pi_k[) = P([0, \pi_k[) = \pi_k$$

Et les tirages sont indépendants.

Sélection des unités

2. Tirage aléatoire simple stratifié (TASST)

A l'intérieur d'une strate, sélection des n unités k ayant les n plus petits numéros aléatoires $g_k(\omega_k)$.

Et comme les nombres tranformés $g_k(\omega_k)$ sont indépendants et suivent la loi uniforme P…

→ Les n plus petits nombres $g_k(\omega_k)$ fournissent donc bien un échantillon aléatoire simple de taille n dans la strate considérée.

$$p(S) = \binom{N}{n}^{-1}$$ (N : taille de la population)

Une procédure “pas à pas” pour prendre en compte les charges de réponse

$\boldsymbol{\omega} = (\omega_1,\ldots,\omega_k,\ldots,\omega_N)$ = vecteur des numéros aléatoires permanents attribués aux unités de la population U.

$I_{k,t}(\boldsymbol{\omega})$ = Indicatrice, égale à 1 si l’unité k est sélectionnée dans l’échantillon t, et 0 sinon : $k \in S_t \Leftrightarrow I_{k,t}(\boldsymbol{\omega}) = 1$

$\gamma_{k,t}$ = charge de réponse pour l’entreprise k lors de l’enquête t.
La charge réelle de réponse est aléatoire : $\gamma_{k,t}(\boldsymbol{\omega}) = \gamma_{k,t}\, I_{k,t}(\boldsymbol{\omega})$

Charge cumulée pour l’unité k : $\Gamma_{k,t}(\boldsymbol{\omega}) = \sum_{u \le t} \gamma_{k,u}.I_{k,u}(\boldsymbol{\omega})$

Principe : définir des fonctions $g_{k,t}$ de sorte que :
$\Gamma_{k,t-1}(\boldsymbol{\omega}_{(1)}) < \Gamma_{k,t-1}(\boldsymbol{\omega}_{(2)}) \Rightarrow g_{k,t}(\omega_{k,(1)}) < g_{k,t}(\omega_{k,(2)})$

Deux difficultés

1. Substituer à $\Gamma_{k,t}(\boldsymbol{\omega})$ une fonction de ω_k seulement, notée $\Gamma'_{k,t}(\omega_k)$, qui founit une bonne approximation de $\Gamma_{k,t}(\boldsymbol{\omega})$.

Tirage de Poisson : $I_{k,t}(\boldsymbol{\omega})$ dépend seulement de ω_k ($I_{k,t}(\boldsymbol{\omega}) = I_{k,t}(\omega_k)$ = 1 si $g_{k,t}(\omega_k) \in [0, \pi_k[$, 0 sinon)
$\rightarrow$ $\Gamma_{k,t}(\boldsymbol{\omega})$ dépend seulement de ω_k , noté $\Gamma'_{k,t}(\omega_k)$.

TASST: $I_{k,t}(\boldsymbol{\omega})$ dépend "essentiellement" de ω_k, mais pas seulement… On approxime $I_{k,t}(\boldsymbol{\omega})$ par $I'_{k,t}(\boldsymbol{\omega}) = E(I_{k,t}(\boldsymbol{\omega})| \omega_k)$, puis $\Gamma_{k,t}(\boldsymbol{\omega})$ par $\Gamma'_{k,t}(\omega_k)$.

2. Définir la fonction de coordination $g_{k,t}$ de sorte que :

$$\Gamma'_{k,t-1}(\omega_{k,1}) < \Gamma'_{k,t-1}(\omega_{k,2}) \Rightarrow g_{k,t}(\omega_{k,1}) < g_{k,t}(\omega_{k,2})$$

Construction d'une fonction de coordination

Pour un critère général de coût $C(\omega_k)$, comme la charge de réponse cumulée approximée : plus il est petit, plus on souhaite que la probabilité de sélection de l'unité k dans l'échantillon t soit élevée.

$$C(\omega_1) < C(\omega_2) \Rightarrow g_C(\omega_1) < g_C(\omega_2) \quad (1)$$

Si C borélienne, on peut construire une première fonction G_C :

$$G_C(\omega) = P^C\left(]-\infty, C(\omega)[\right) = P\left(C^{-1}]-\infty, C(\omega)[\right) = P\left(u \middle| C(u) < C(\omega)\right)$$

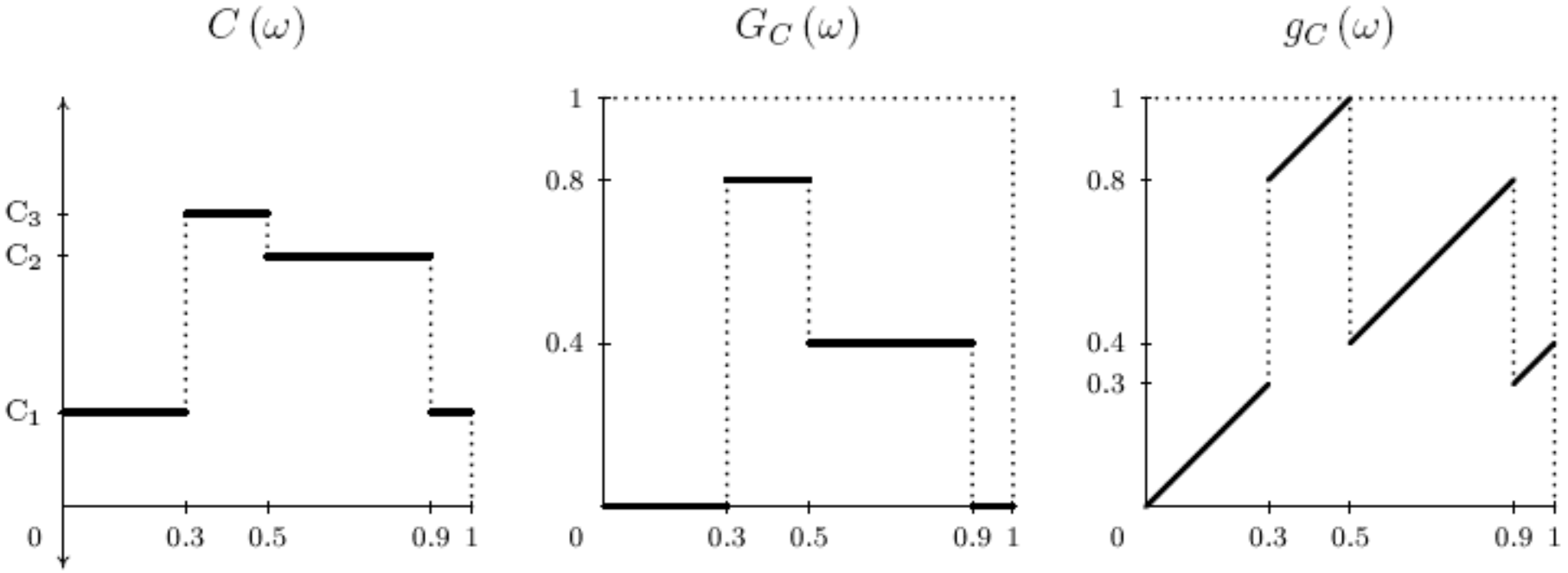
$$"G_C = F_C(C)"$$

Si C admet des paliers (boréliens de [0;1[de mesure de Lebesgue non nulle sur lesquels C est constante), alors G_C admet des paliers et ce sont les mêmes que pour C.

Si pas de palier, G_C est une fonction de coordination vérifiant (1) !!!

Construction d'une fonction de coordination

Et si C, donc G_C, ont des paliers ? Une simple transformation affine des paliers permet de modifier G_C en une fonction de coordination g_C satisfaisant (1) !!!



On peut ensuite introduire des critères de coûts secondaires…

Application au Tirage de Poisson

Initialisation : $\Gamma_{k,0}(\omega_k) = 0 \quad g_{k,1}(\omega_k) = \omega_k$

$$I_{k,1}(\omega_k) = \mathbb{1}_{[0,\pi_{k,1}[}(\omega_k) \quad \Gamma_{k,1}(\omega_k) = \gamma_{k,1}\,\mathbb{1}_{[0,\pi_{k,1}[}(\omega_k)$$

Pour l'échantillon S_t, on choisit une fonction de coordination $g_{k,t}$ associée à l'unité k.

$$k \in S_t \Leftrightarrow g_{k,t}(\omega_k) \in [0,\pi_{k,t}[$$

Si l'on définit : $A_{k,t} = g_{k,t}^{-1}[0,\pi_{k,t}[$

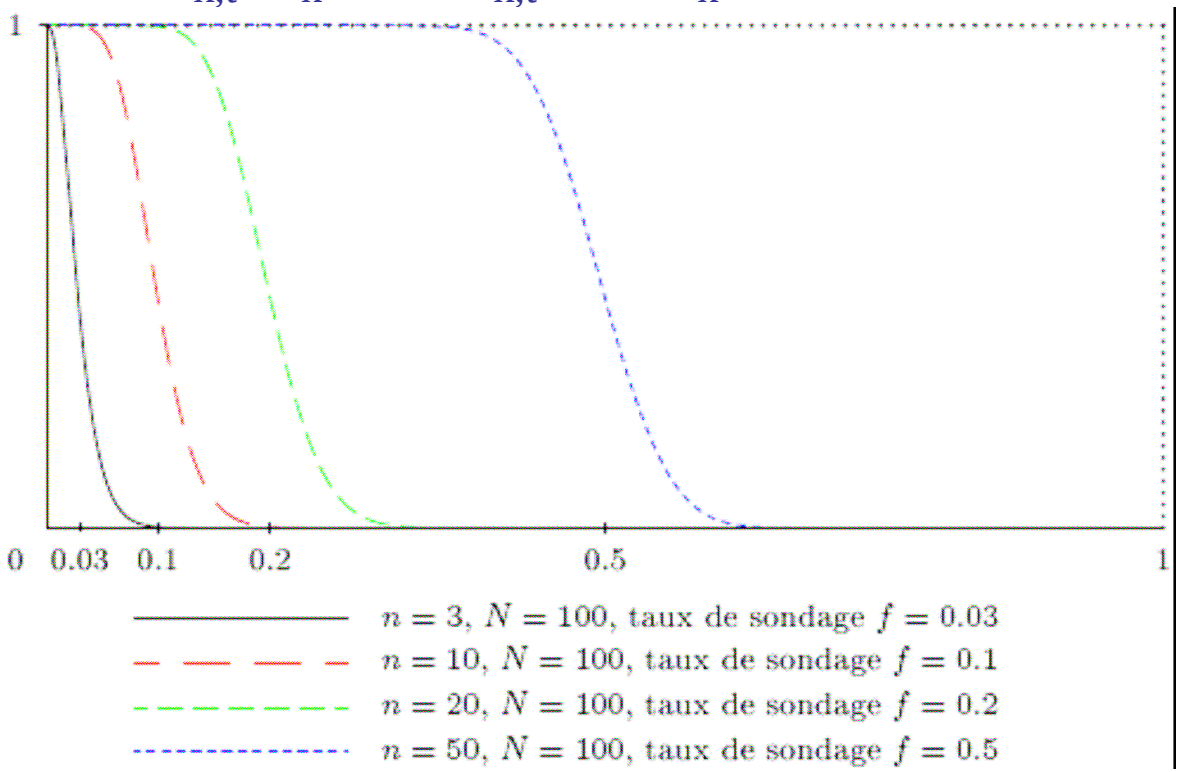
$\rightarrow$indicatrice de sélection : $I_{k,t}(\omega_k) = \mathbb{1}_{A_{k,t}}(\omega_k)$

Pour la sélection d'un échantillon « séparé » S_t :

critère $C_{k,t}(\omega_k)$ = charge cumulée $\Gamma_{k,t-1}(\omega_k)$, on en déduit $g_{k,t}$

Application au TASST

1. Approximation $I'_{k,t}(\omega_k) = E(I_{k,t}(\boldsymbol{\omega})|\ \omega_k)$,



2. Implémentation informatique : $I'_{k,t}$ et charges cumulées espérées ***approximées*** par des fonctions constantes par morceaux ([0;1[subdivisé en L sous-intervalles). → Puis procédure similaire au Tirage de Poisson

Première simulation, cas simple

Population fixe dans le temps de taille 100.

Charge initiale = 0 pour chaque unité.

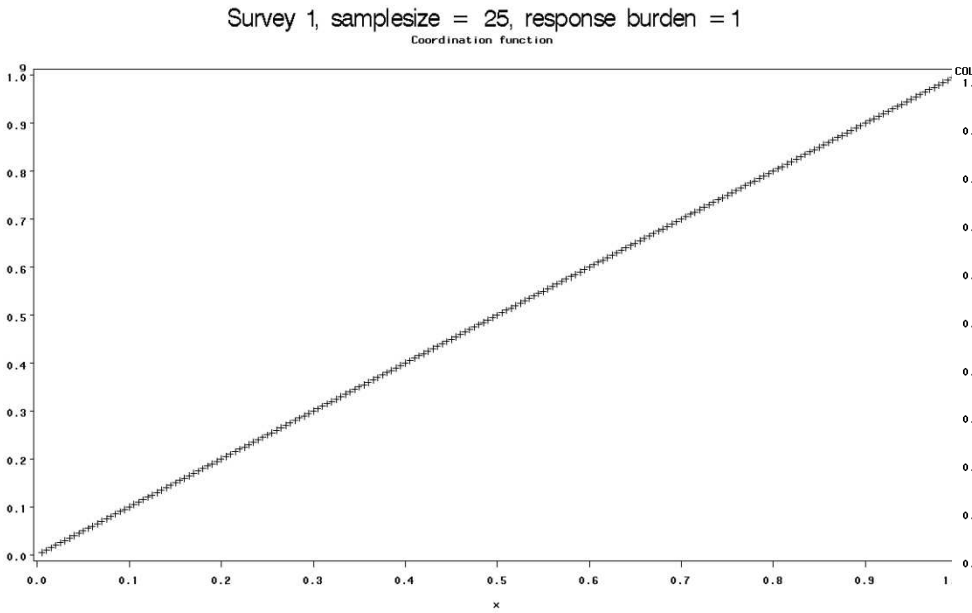
Sélection de 20 échantillons (SAS).

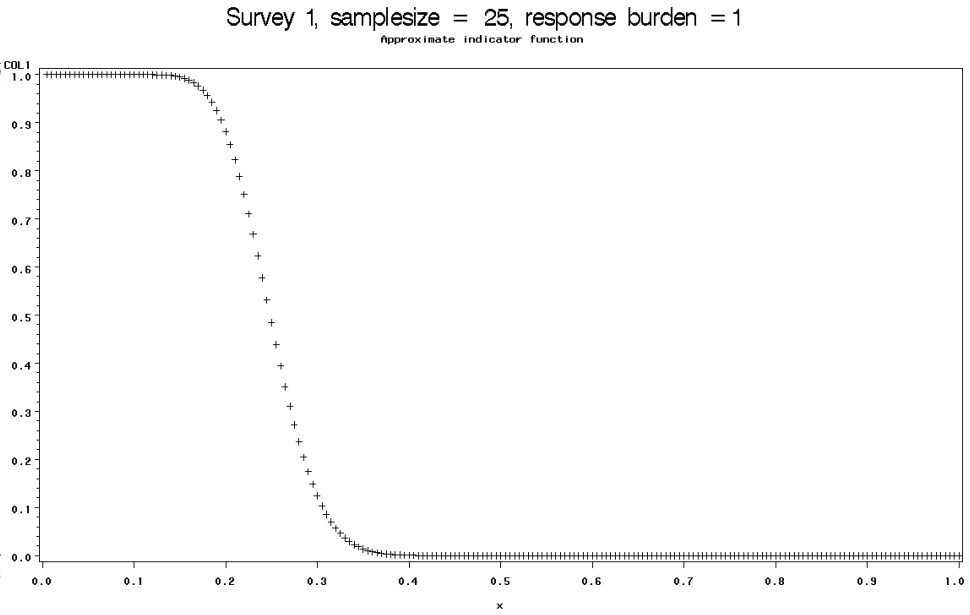
Pour chaque échantillon : taille n = 25, charge de réponse = 1,

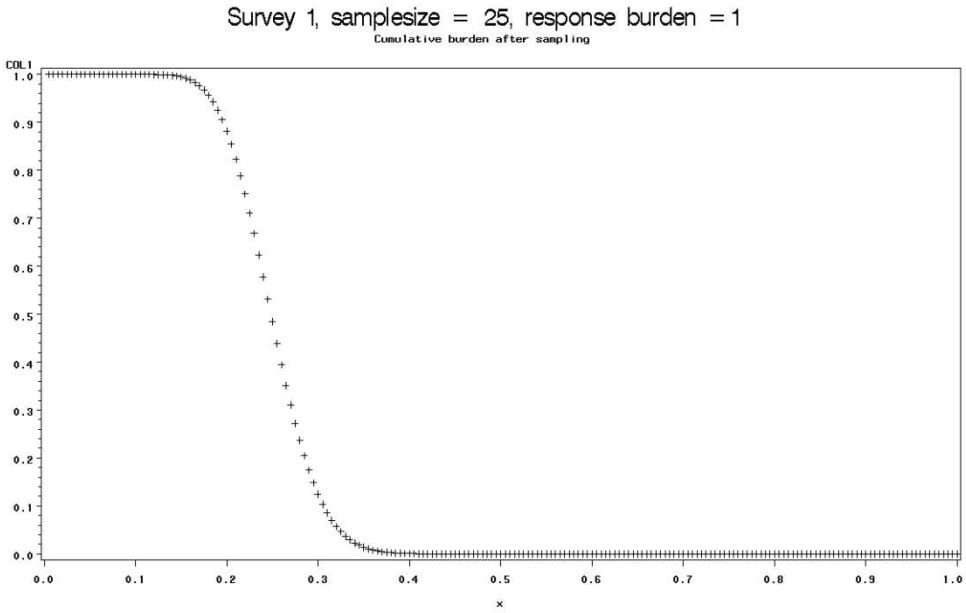
sauf :

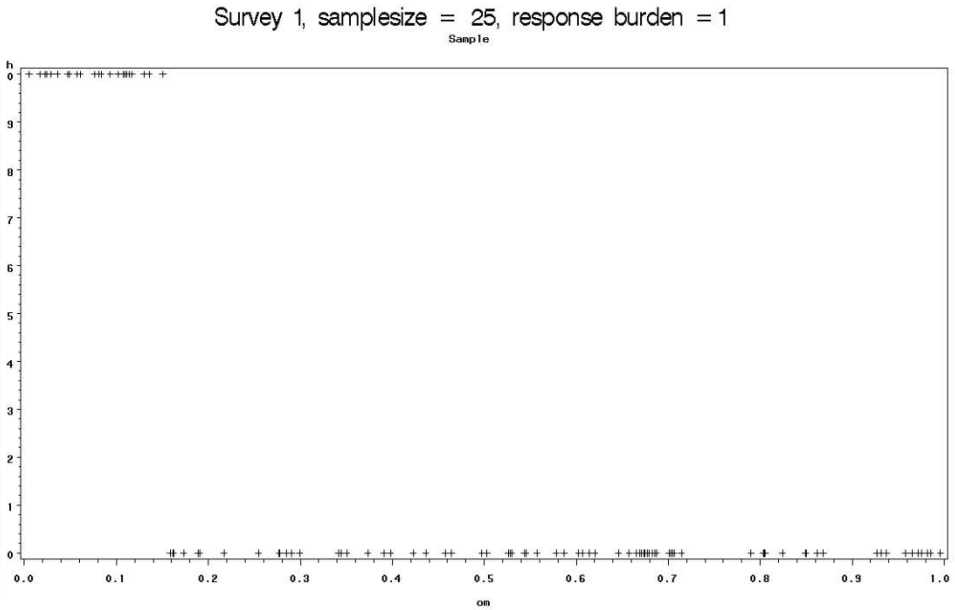
Echantillons n°3 et n°15 : taille n = 50, charge de réponse = 3
Echantillons n°10 et n°11 : taille n = 10, charge de réponse = 2

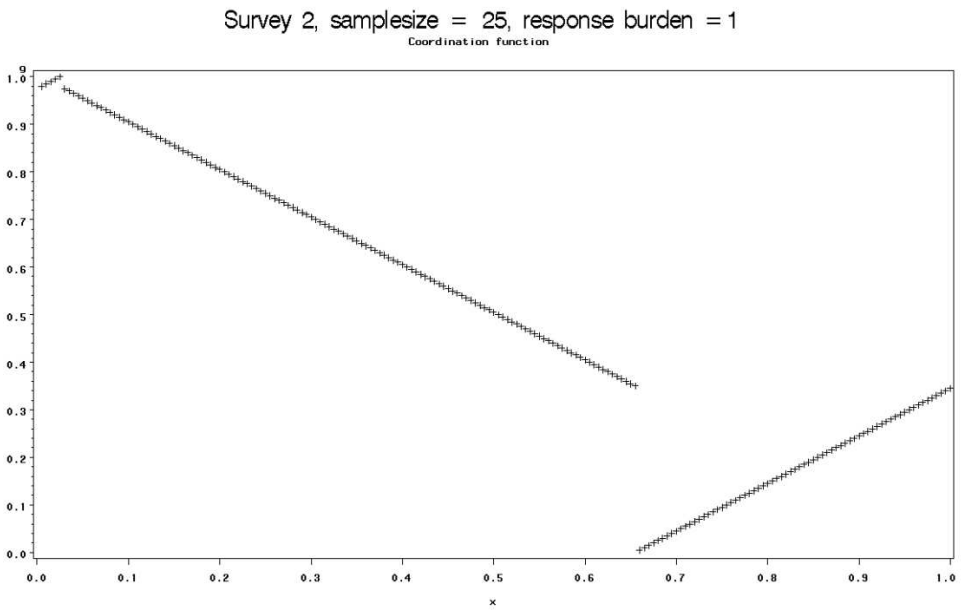
→ Charge cumulée moyenne espérée = 7.4

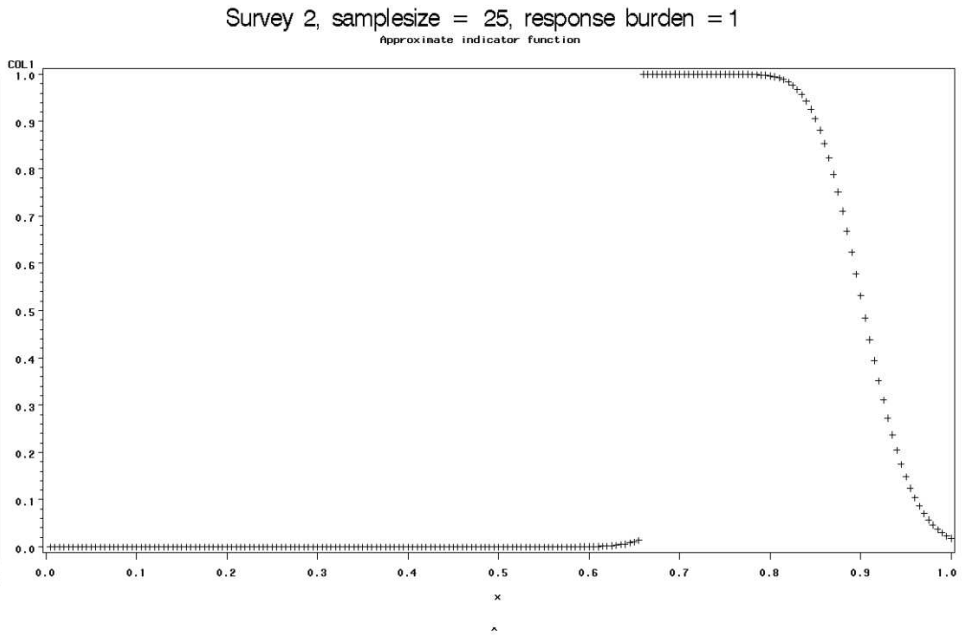

Survey 1, samplesize = 25, response burden = 1
Coordination function
x

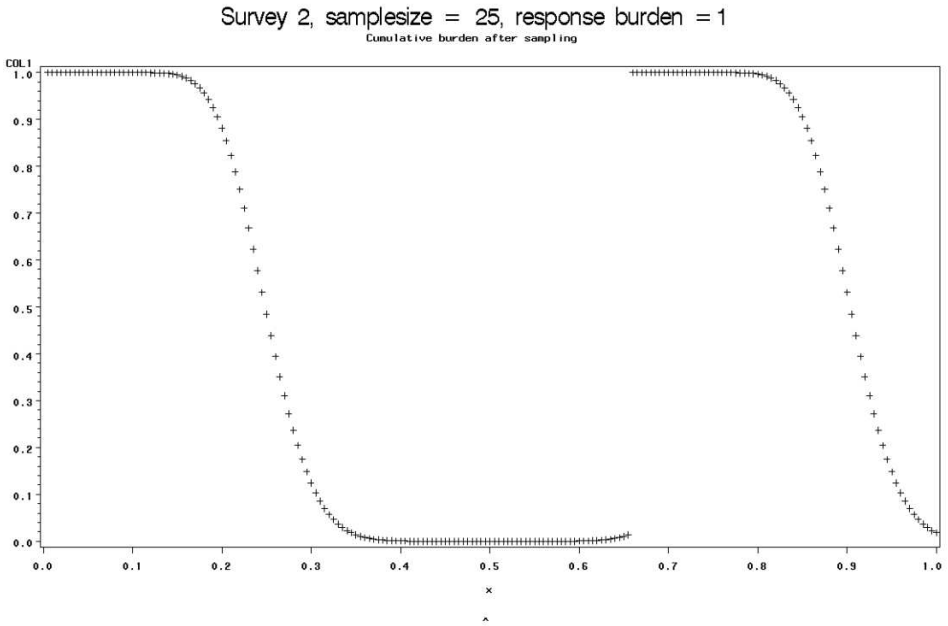

Survey 1, samplesize = 25, response burden = 1
Approximate indicator function
COL1
x

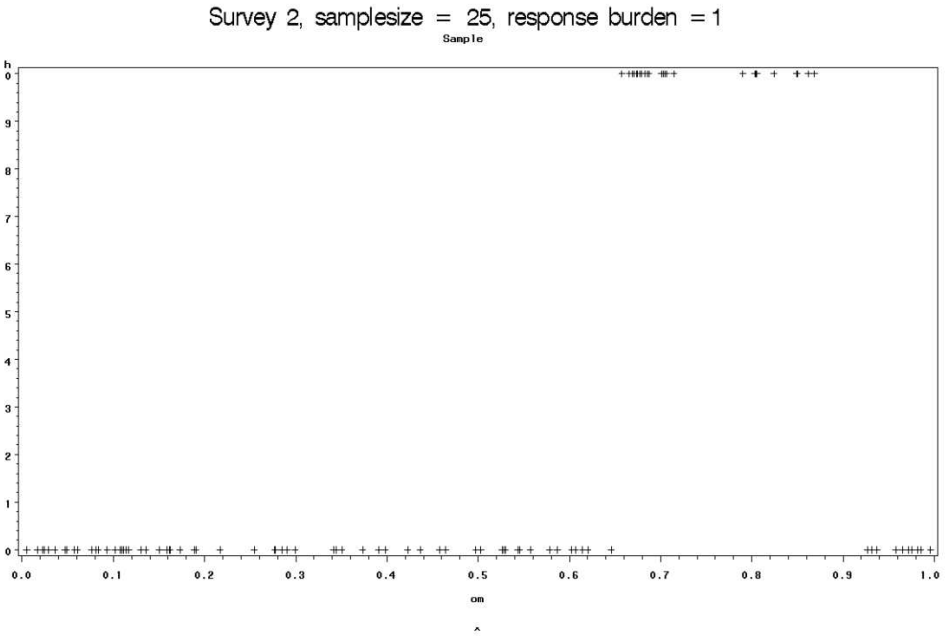

Survey 1, samplesize = 25, response burden = 1
Cumulative burden after sampling
COL1
x

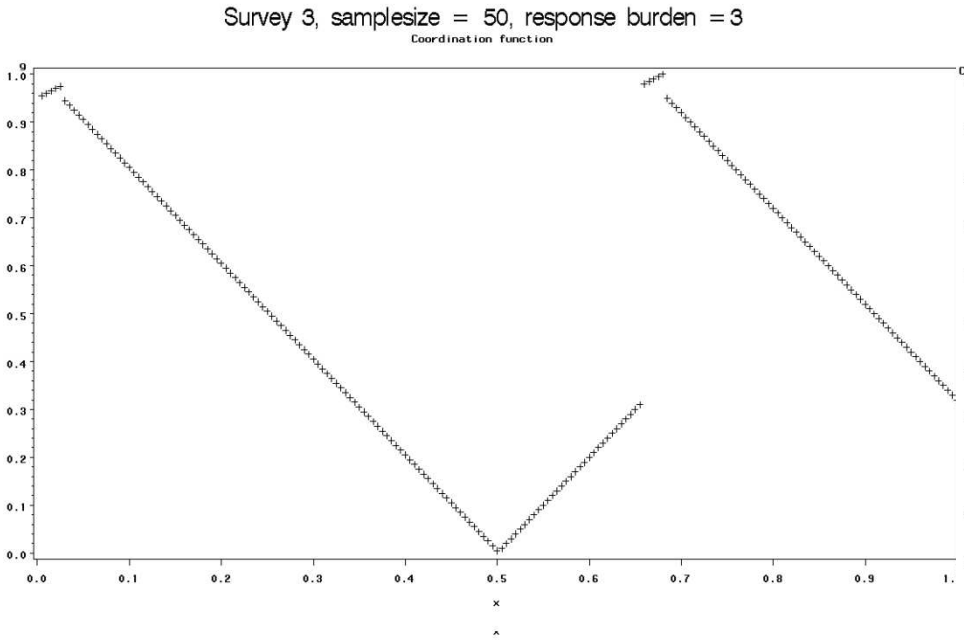

Survey 1, samplesize = 25, response burden = 1
Sample
h
om

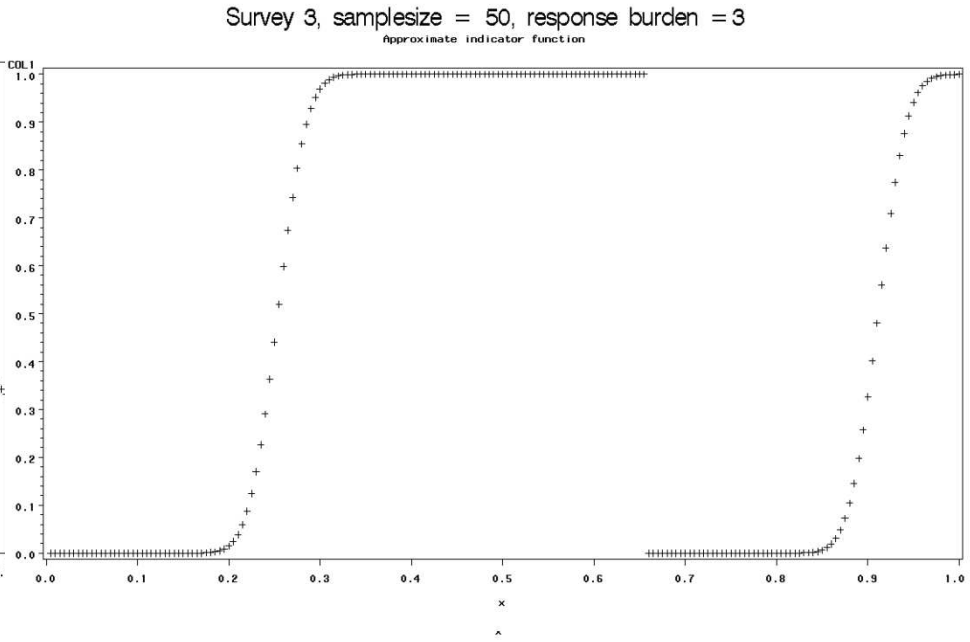

Survey 2, samplesize = 25, response burden = 1
Coordination function
x

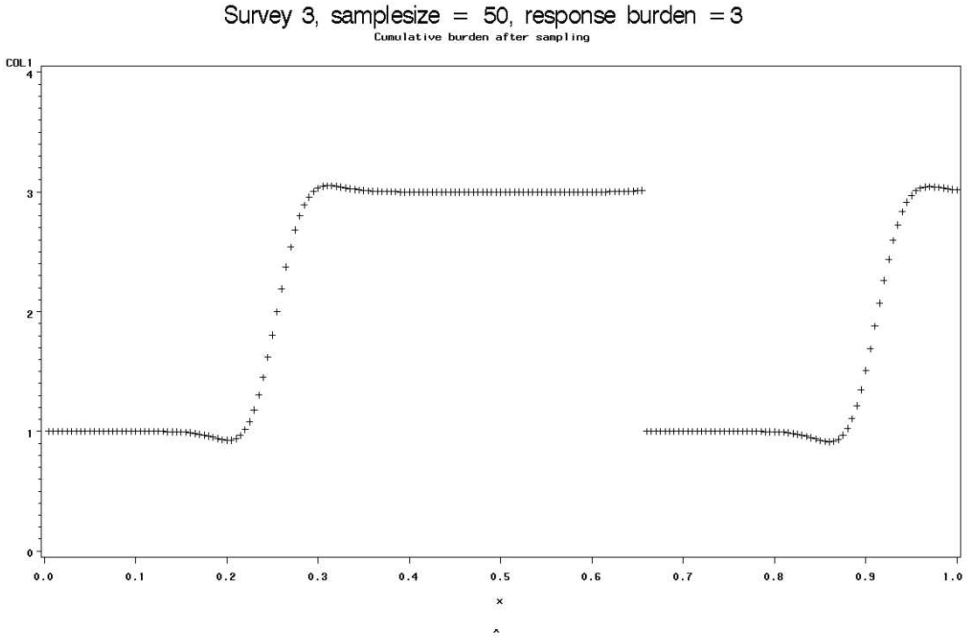

Survey 2, samplesize = 25, response burden = 1
Approximate indicator function
x

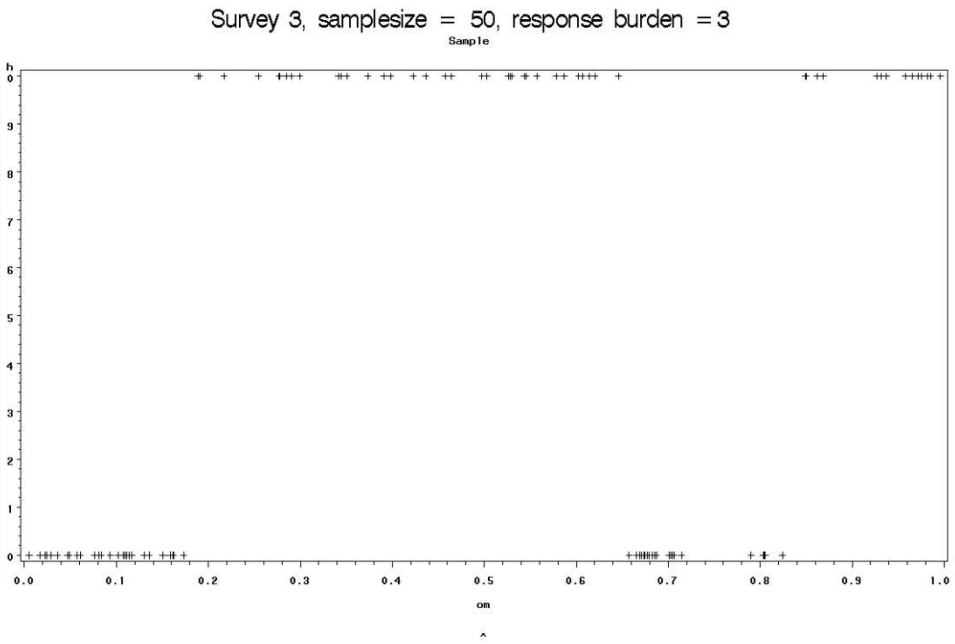

Survey 2, samplesize = 25, response burden = 1
Cumulative burden after sampling
x

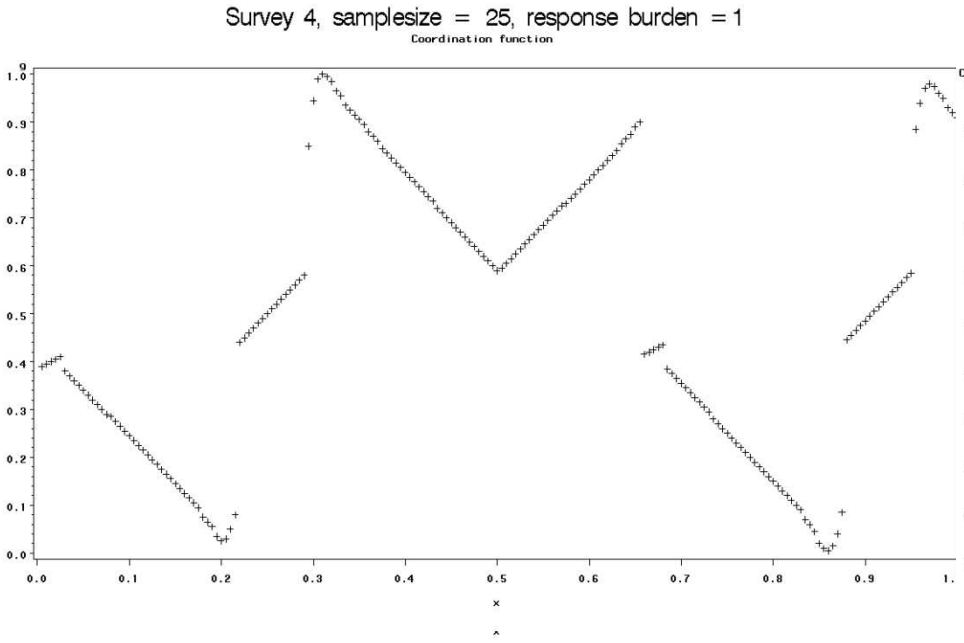

Survey 2, samplesize = 25, response burden = 1
Sample
om

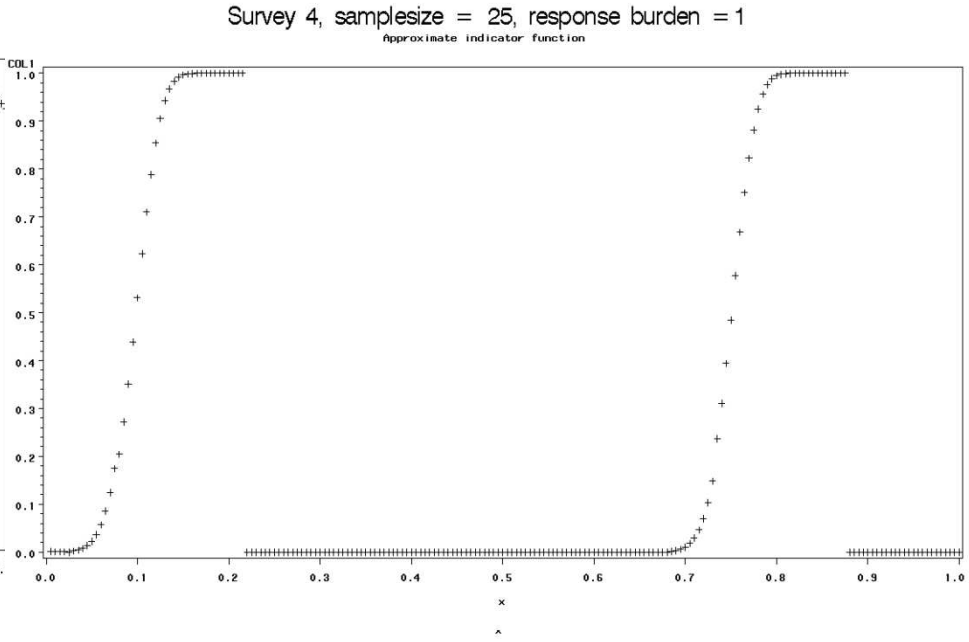

Survey 3, samplesize = 50, response burden = 3
Coordination function
x

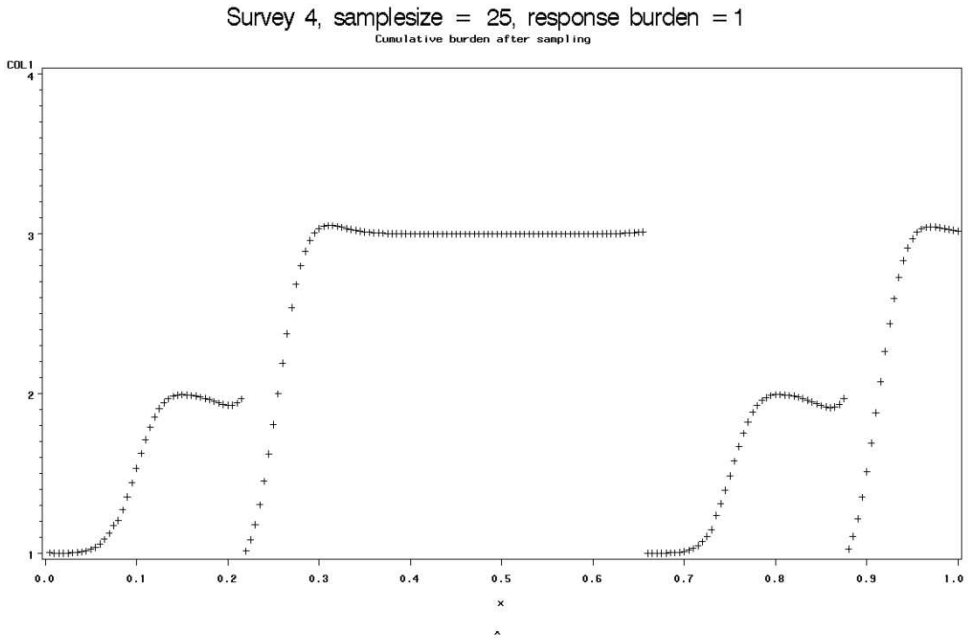

Survey 3, samplesize = 50, response burden = 3
Approximate indicator function
x

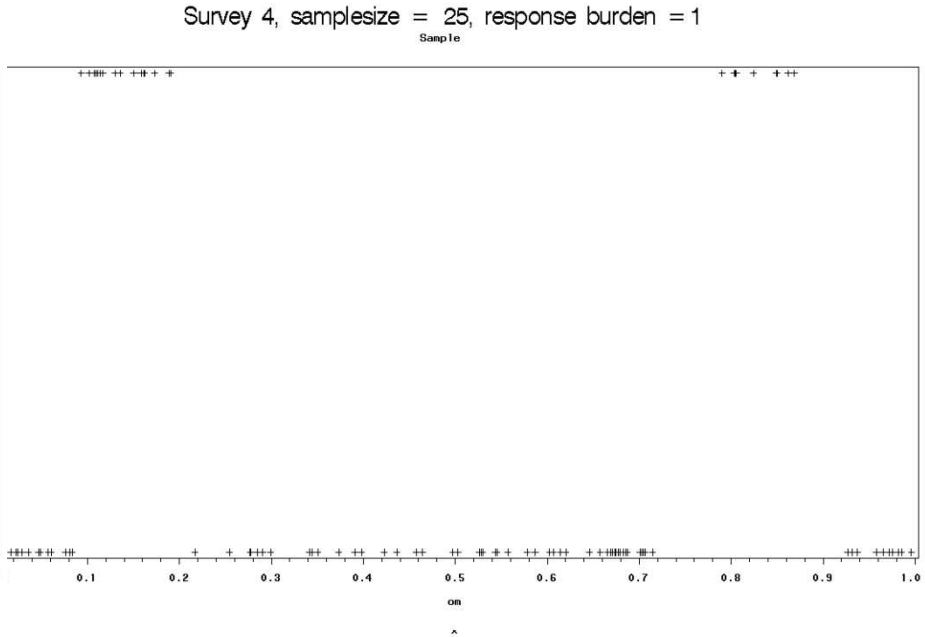

Survey 3, samplesize = 50, response burden = 3
Cumulative burden after sampling
x

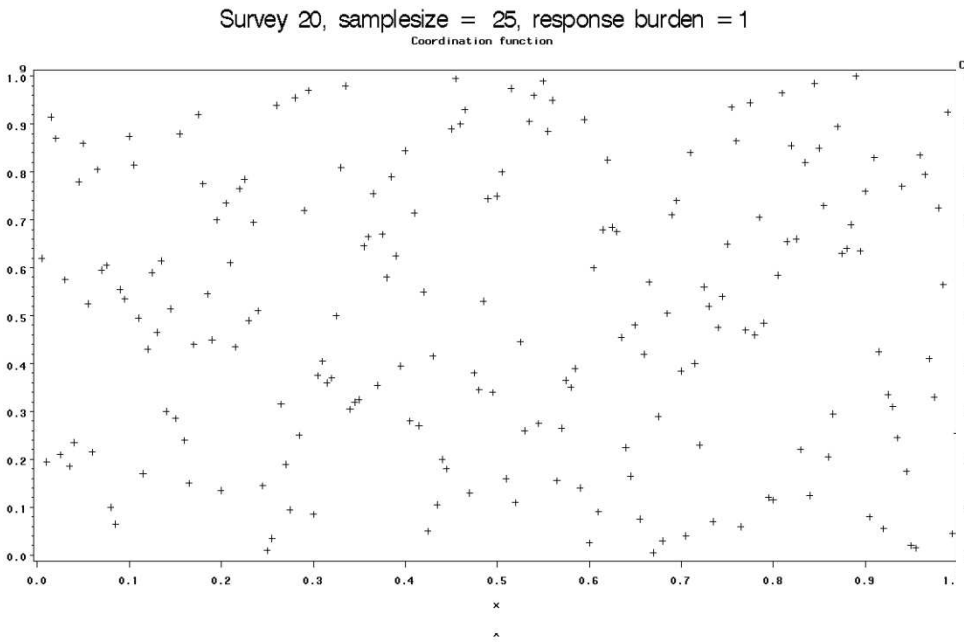

Survey 3, samplesize = 50, response burden = 3
Sample
om

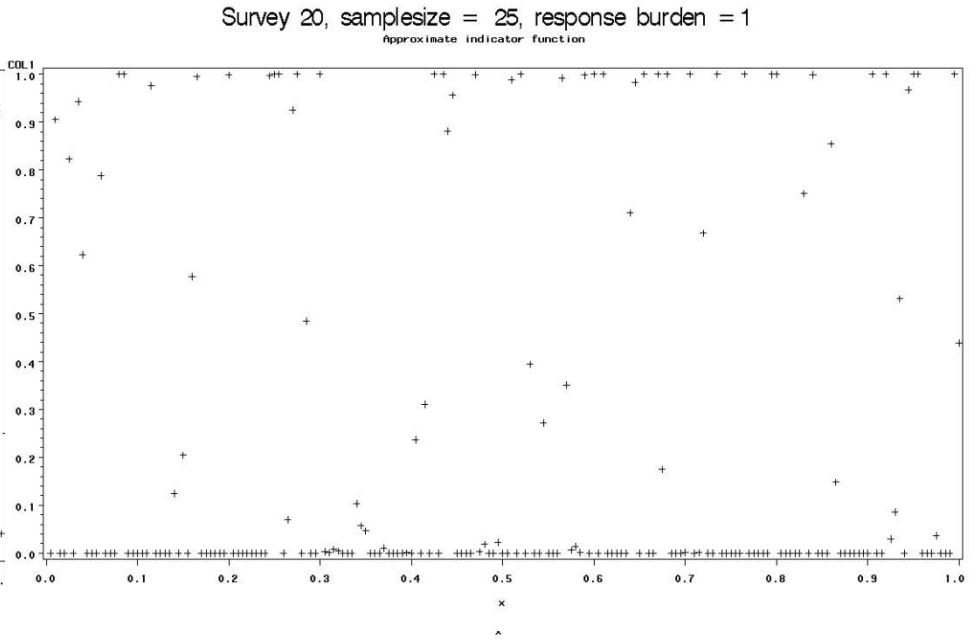

Survey 4, samplesize = 25, response burden = 1
Coordination function
x

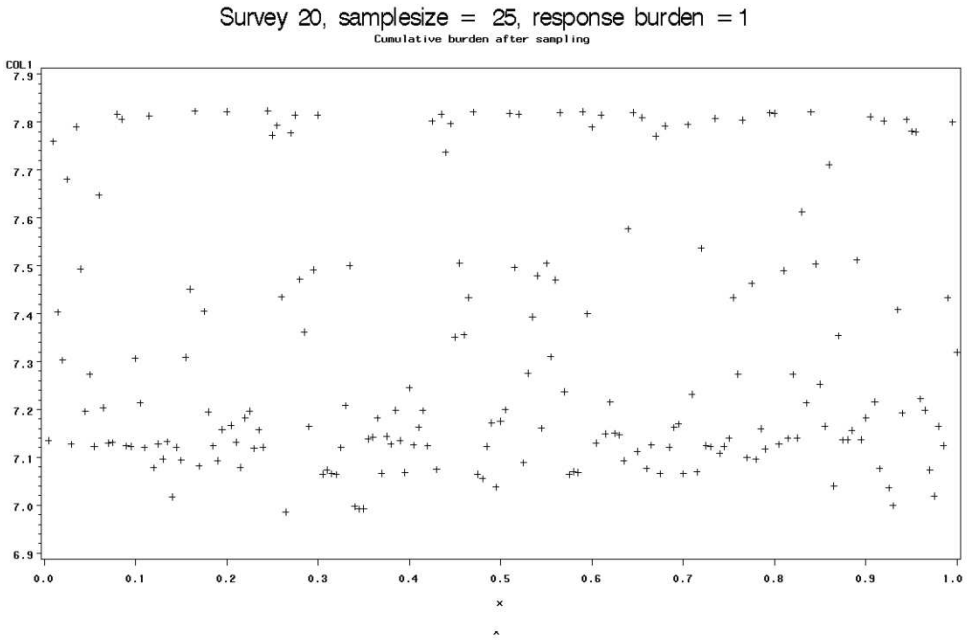

Survey 4, samplesize = 25, response burden = 1
Approximate indicator function
COL1
x

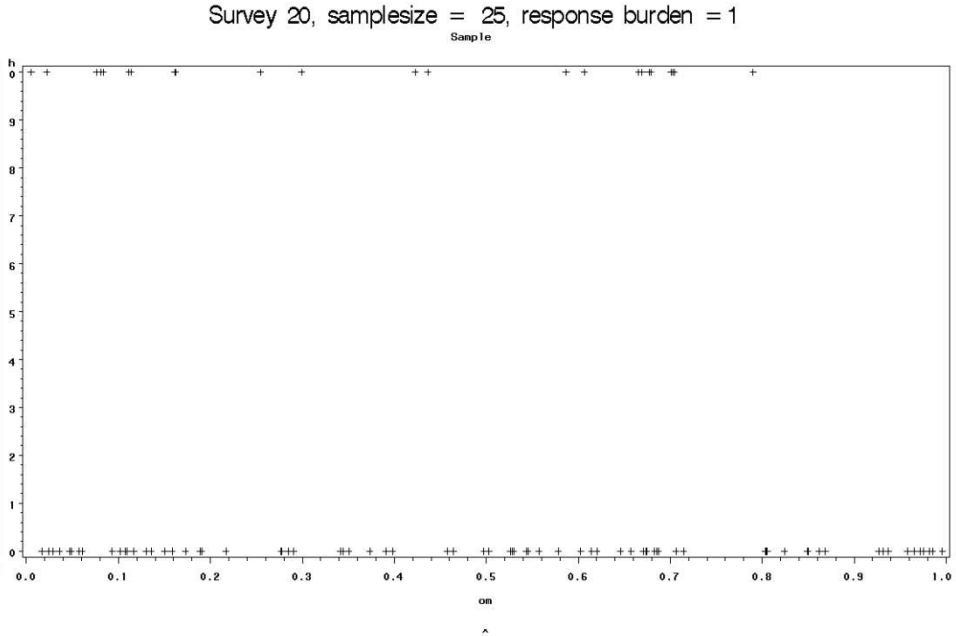

Survey 4, samplesize = 25, response burden = 1
Cumulative burden after sampling
COL1
x


Survey 4, samplesize = 25, response burden = 1
Sample
om


Survey 20, samplesize = 25, response burden = 1
Coordination function
x


Survey 20, samplesize = 25, response burden = 1
Approximate indicator function
x

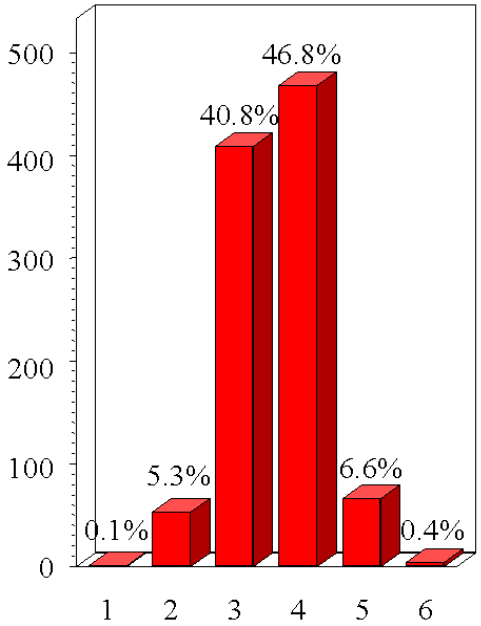

Survey 20, samplesize = 25, response burden = 1
Cumulative burden after sampling
x


Survey 20, samplesize = 25, response burden = 1
Sample
om

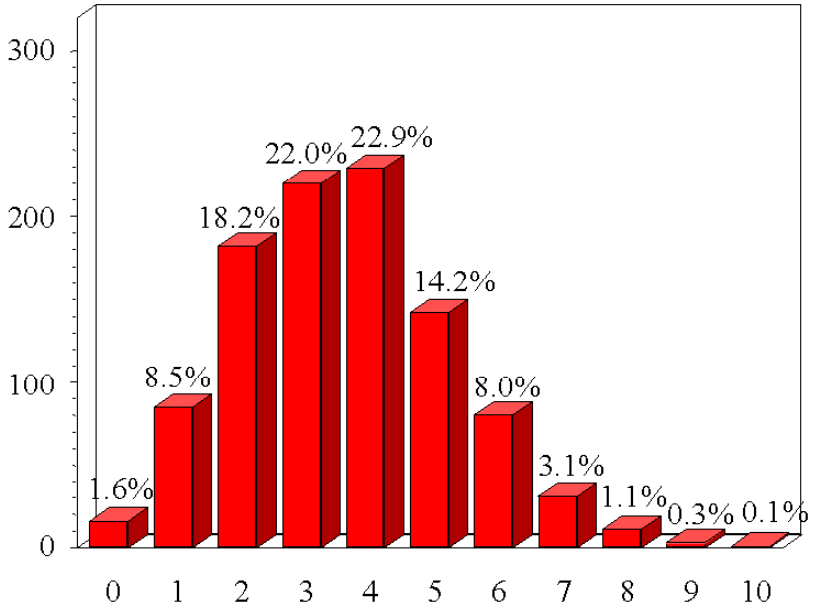
Simulations : performances de la coordination en regard de divers paramètres

Tirage de 20 échantillons successifs (TASST) sur une population évoluant dans le temps.

Nombre d'unités en fct de la fréquence d'échantillonnage



Cas coordonné



Cas indépendant

2 critères de performance : $R_\sigma = \frac{\sigma_{Coord}}{\sigma_{Indep}} < 1$ $\Delta_s = S_{Coord} - S_{Indep} < 0$

Simulations : performances de la coordination en regard de divers paramètres

- Globalement, résultats très positifs : pour quasi toutes les simulations envisagées, $\mathbf{R_\sigma < 0.5}$ (Plus de 50% de gains en termes de répartition de la charge) et $\mathbf{\Delta_S < 0}$

- Robustesse de la méthode vis-à-vis de situations instables : population mouvante, stratif. et taux de sondage très variables d'un tirage à l'autre…

- Importance de coordonner négativement une enquête avec suffisamment d'enquêtes du passé : avec au moins Ncoord enquêtes, où $\mathrm{Ncoord} > \bar{f}^{-1} - 1$ ($\bar{f}$: taux moyen de sondage)

Beaucoup plus de résultats dans l'article associé à cette présentation …

Merci de votre attention !

Contact
M. Fabien Guggemos
Tél. : 00 33 1 41 17 50 18
Mail : fabien.guggemos@insee.fr

M. Olivier Sautory
Tél. : 00 33 1 41 17 50 82
Mail : olivier.sautory@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr

Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00