European Establishment Statistics Workshop

12 – 14 september 2011 - Neuchâtel

Sampling coordination of business surveys conducted by INSEE

F. Guggemos, O. Sautory (Insee)

Random numbers

Very often national statistics agencies use the permanent random numbers technique for the sampling coordination of business surveys.

Each unit k of the population (included the new units) is independently assigned a number ω_k, selected according to the uniform distribution in the interval [0,1 [.

Poisson sampling : selection of the units k whose numbers belong to the interval [d, d + π_k[, π_k = probability of inclusion of the unit k

Simple random sampling (SRS) : selection of the n units with the lowest numbers ω_k superior to d.

Constant shift method : J panels (j = 1 …J)

Starting point for panel j at the date a : $d_{j,a} = d_{j,1} + (a-1)c, \quad a \geq 1$

Definition of a coordination function

Coordination function g = measurable application from [0,1[to [0,1[which preserves uniform probability :

if P is the uniform probability on [0,1[, then the image probability

$$P^g = P.$$

$\rightarrow$ for any interval I = [a,b [included in [0,1[:

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

Selection of the units

For each unit k in the sampling frame :

- a permanent random number ω_k,
- a coordination function $g_{k,t}$ that changes at each sampling t = 1, 2, …

1. Poisson sampling

Selection of the units k such that $g_k(\omega_k) \in [0, \pi_k[$
where π_k = probability of inclusion of the unit k

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

and the drawings are independent.

Selection of the units

2. SSRS

Within a stratum, selection of the n units k associated with the n smallest numbers $g_k(\omega_k)$.

Since

- P^{g_k} = uniform probability P sur [0,1[for each k

- and the n numbers (ω_k) are drawn independently from P

→ the n numbers $g_k(\omega_k)$ are drawn independently from P

→ the n smallest numbers $g_k(\omega_k)$ give a simple random sample of size n in the stratum.

Example : the constant shift method

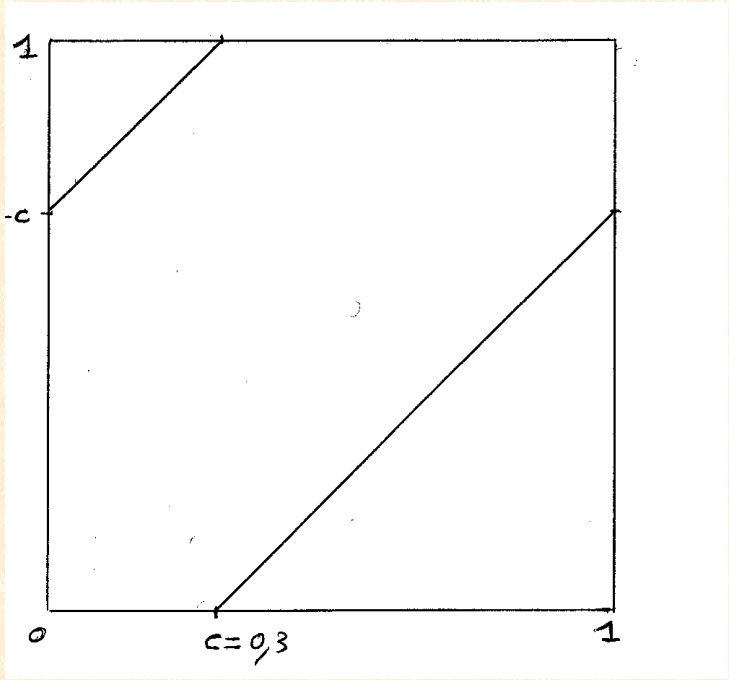
Let $d_1 = 0$ and $d_2 = c$. We define the coordination functions :

$$g_{k,1}(\omega_k) = \omega_k \qquad g_{k,2}(\omega_k) = \omega_k - c \pmod 1$$

Then :

$$k \in S_1 \Leftrightarrow \omega_k \in [0, \pi_{k,1}[\Leftrightarrow g_{k,1}(\omega_k) \in [0, \pi_{k,1}[$$

$$k \in S_2 \Leftrightarrow \omega_k \in [c, c + \pi_{k,2}[\Leftrightarrow g_{k,2}(\omega_k) \in [0, \pi_{k,2}[$$



A step by step procedure reflecting response burdens

$\boldsymbol{\omega} = (\ldots\ \omega_k \ldots)$ = vector of random numbers given to the population units.

$I_{k,t}(\boldsymbol{\omega})$ = indicator function, equal to 1 if the values in $\boldsymbol{\omega}$ lead to select the unit k in the sampling t, and 0 otherwise :

$$k \in S_t \Leftrightarrow I_{k,t}(\boldsymbol{\omega}) = 1$$

$\gamma_{k,t}$ = response burden of a questioned enterprise k at survey t

Effective burden = random variable $\gamma_{k,t}(\boldsymbol{\omega}) = \gamma_{k,t}\ I_{k,t}(\boldsymbol{\omega})$

Cumulative burden for unit k : $\Gamma_{k,t}(\boldsymbol{\omega}) = \sum_{u \le t} \gamma_{k,u} . I_{k,u}(\boldsymbol{\omega})$

<u>Principle</u> : to define the coordination functions $g_{k,t}$ for the selection of sample S_t using $\Gamma_{k,t-1}$:

$$\Gamma_{k,t-1}(\boldsymbol{\omega_1}) < \Gamma_{k,t-1}(\boldsymbol{\omega_2}) \Rightarrow g_{k,t}(\omega_{k,1}) < g_{k,t}(\omega_{k,2})$$

Two difficulties

1. to substitute for $\Gamma_{k,t}(\boldsymbol{\omega})$ a function of ω_k only, denoted $\Gamma'_{k,t}(\omega_k)$, that closely approximates $\Gamma_{k,t}(\boldsymbol{\omega})$.

Poisson sampling : $I_{k,t}(\boldsymbol{\omega})$ depends only on ω_k (indicator function of an interval of length π_k), and can be denoted $I_{k,t}(\omega_k)$
$\rightarrow$ $\Gamma_{k,t}(\boldsymbol{\omega})$ depends only on ω_k , and can be denoted $\Gamma'_{k,t}(\omega_k)$.

SSRS : $I_{k,t}(\boldsymbol{\omega})$ depends on all coordinates of the vector $\boldsymbol{\omega}$, but "primarily" on coordinate ω_k : if we select the $\boldsymbol{n}$ units among $\boldsymbol{N}$ with the $\boldsymbol{n}$ smallest values ω_j, it will be equal to 1 for values of ω_k near to 0, regardless of the values of the other coordinates $\rightarrow$ we will be able to replace $I_{k,t}(\boldsymbol{\omega})$ with an approximation $I'_{k,t}(\omega_k)$, and therefore to replace $\Gamma_{k,t}(\boldsymbol{\omega})$ with an approximation $\Gamma'_{k,t}(\omega_k)$.

2. to define the coordination function $g_{k,t}$ such that :

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

Construction of a coordination function

$C_{k,t}(\omega_k)$ = criterion such that the smaller is the criterion, the larger is the probability of selection for unit k at sampling t.

We drop the subscripts k and t. C is supposed to be a bounded measurable function :

$$\omega \in [0,1[\rightarrow C(\omega) \in \mathbb{R}$$

Idea : to associate to this criterion a coordination function g_C such that :

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

P^C = image probability of P under C, F_C the distribution function of C.

The coordination function is built from $G_C = F_C(C)$:

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

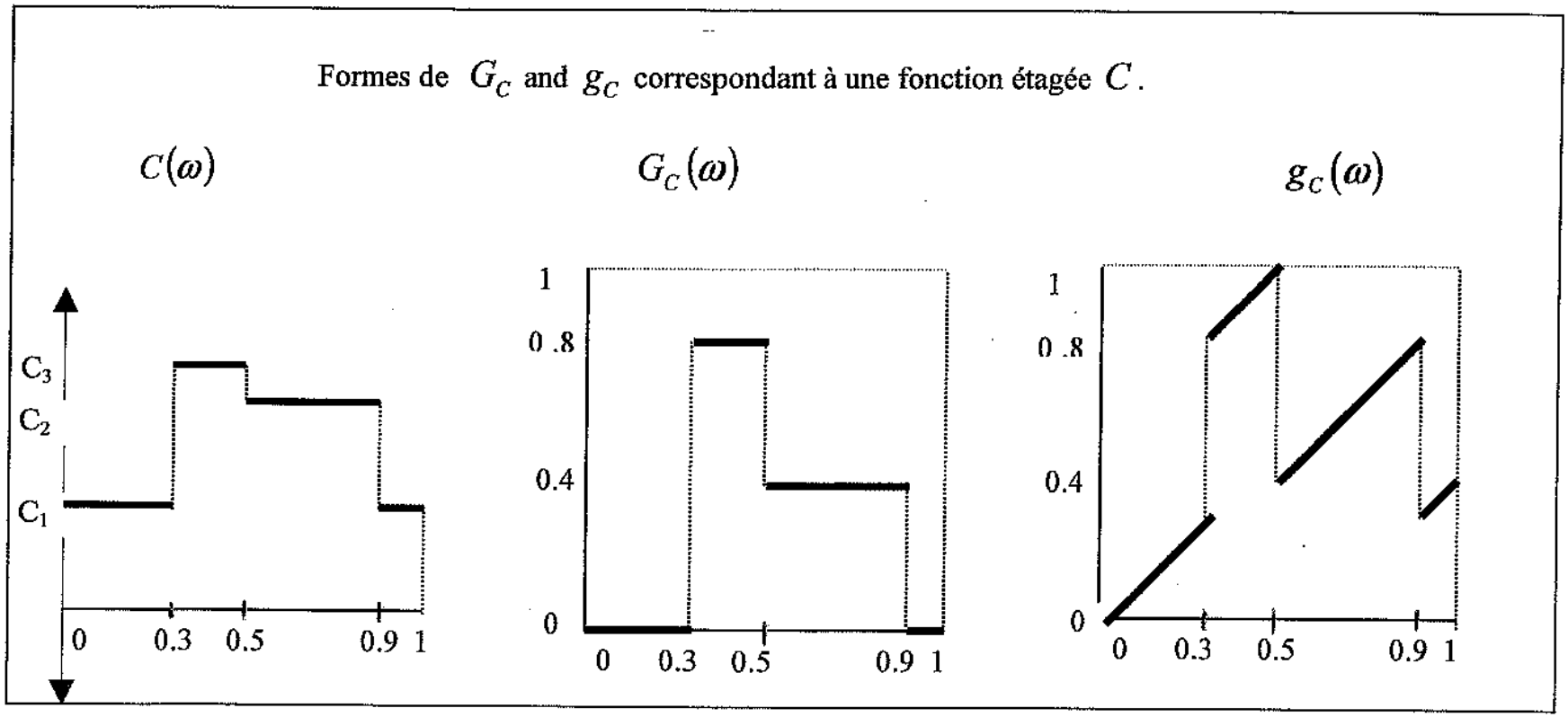
The way to derive g_C from G_C depends on whether or not C has *levels*.

Definition : we call a *level* of criterion C any inverse image of a real number y such that $P^C(y) = P(A) > 0$
i.e. C has *levels* when horizontal line segments form part of the graph of C.

Properties of G_C

- The range of G_C is included in [0,1[
- G_C has the same *levels* as C
- G_C verifies implication (1)
- For every y in the range of G_C, we have $P\left(u \middle| G_C(u) < y\right) = y$
- If C has no *level*, G_C is a coordination function.

If C has at least one level, the range of function G_C is strictly included in [0,1[. We have to deduce from G_C another function, denoted g_C, such that the range of g_C is equal to [0,1[.



Coordination function with several criteria

If criterion C has levels A_i, we can introduce secondary criteria C_i corresponding to each level.

The coordination function verifies the following conditions:

$$\forall\ \omega_1, \omega_2 \in A_i \quad C_i(\omega_1) < C_i(\omega_2) \Rightarrow g(\omega_1) < g(\omega_2)$$

Application to Poisson sampling

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

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

For sample S_t, we choose a coordination function $g_{k,t}$ associated to each unit k. Then :

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

We define : $A_{k,t} = g_{k,t}^{-1}[0,\pi_{k,t}[$

$\rightarrow$ indicator function : $I_{k,t}(\omega_k) = \mathbb{1}_{A_{k,t}}(\omega_k)$

Application to Poisson sampling

Sampling

For a separate sample t, for the selection of the sample S_t :

criterion $C_{k,t}(\omega_k)$ = cumulative burden $\Gamma_{k,t-1}(\omega_k)$, and then deduce $g_{k,t}$

For updating a panel : S_u = sample corresponding to the latest update ($u \leq t-1$). We denote $A_{k,u} = g_{k,u}^{-1}[0, \pi_{k,u}[$

First-stage criterion in the calculation of the coordination function :
any decreasing function of the indicator function of $A_{k,u}$.
For example : $C_{k,t}(\omega_k) = 1$ if $\omega_k \in A_{k,u}$ (i.e. $k \in S_u$)
$= 2$ if $\omega_k \notin A_{k,u}$ (i.e. $k \notin S_u$)

To take into account past burdens :
cumulative burden $\Gamma_{k,t-1}$ = secondary criterion
This leads to a certain coordination function $g_{k,t}$.

Application to SSRS

1. Calculation of the approximate $I'_{k,t}(\omega_k)$, by its conditional expectation

$$I'_{k,t}(\omega_k) = E\left(I_{k,t}(\Omega) \middle| \Omega = \Omega_k\right) = b_{k,t}\left(g_{k,t}(\omega_k)\right)$$

where $\Omega = (\Omega_1 \ldots \Omega_k \ldots \Omega_N)$ is a random vector from which we have a realization ω.

2. Approximation by step functions

$I'_{k,t}$ and the cumulative burden function are no longer step functions. These functions are approximated by step functions, which are constant over predefined intervals, obtained by dividing [0, 1[in L equal subintervals.

Then, the procedures are similar to those used in Poisson sampling.

Simulation

Fixed population of size 100.

Initial burden = 0 for each unit.

Selection of 20 samples (simple random sampling).

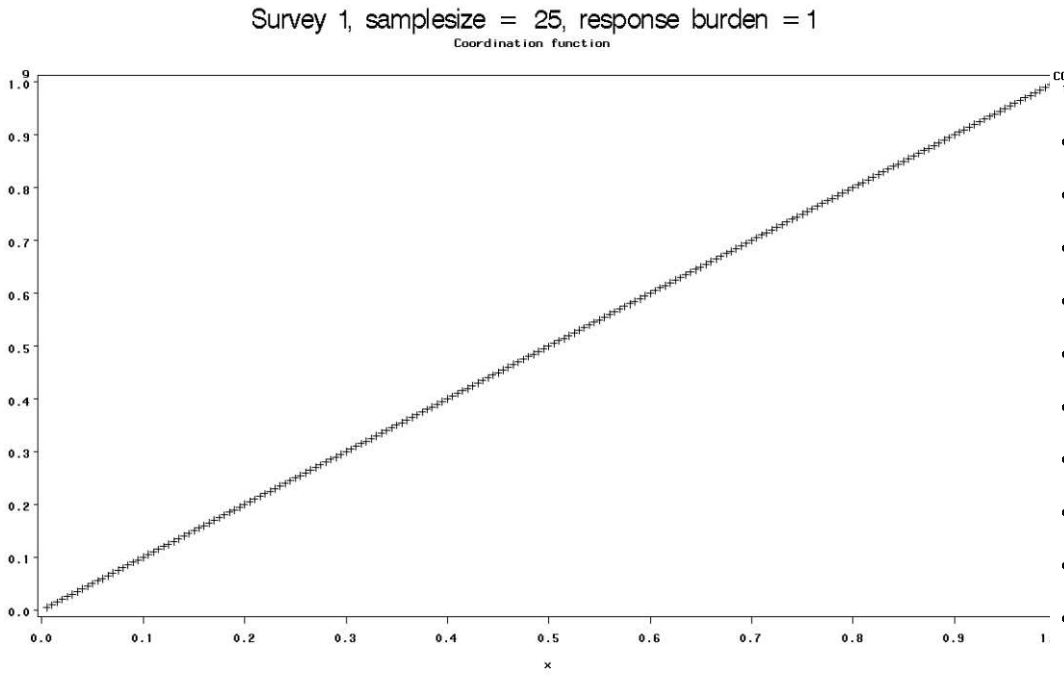
For each sample : sample size = 25, response burden =1,

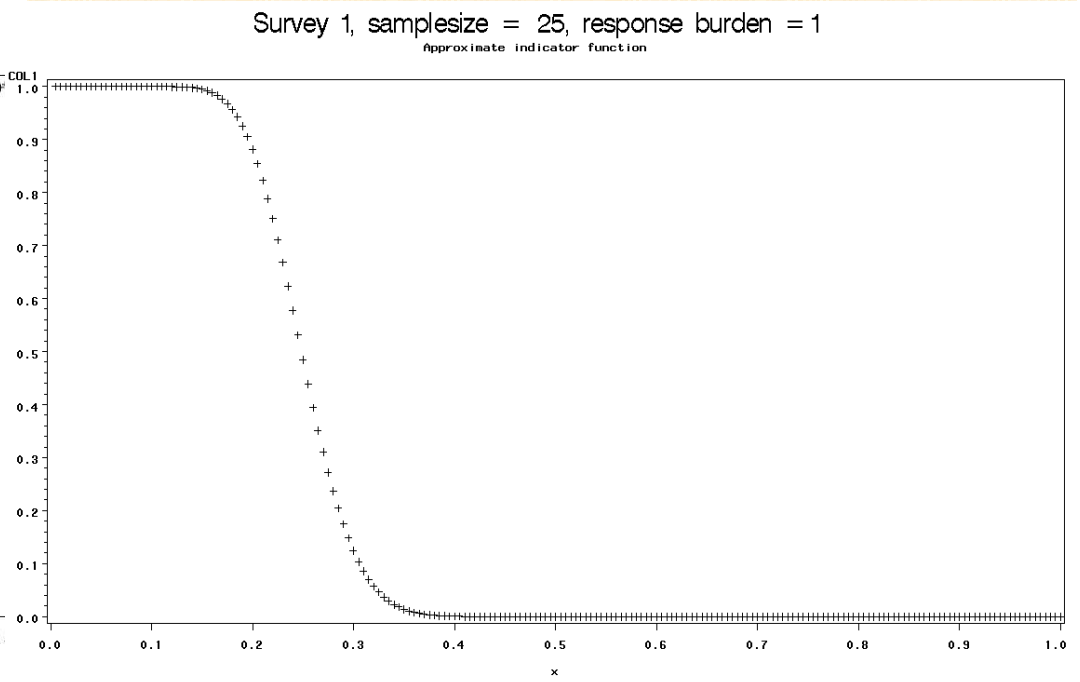
except :

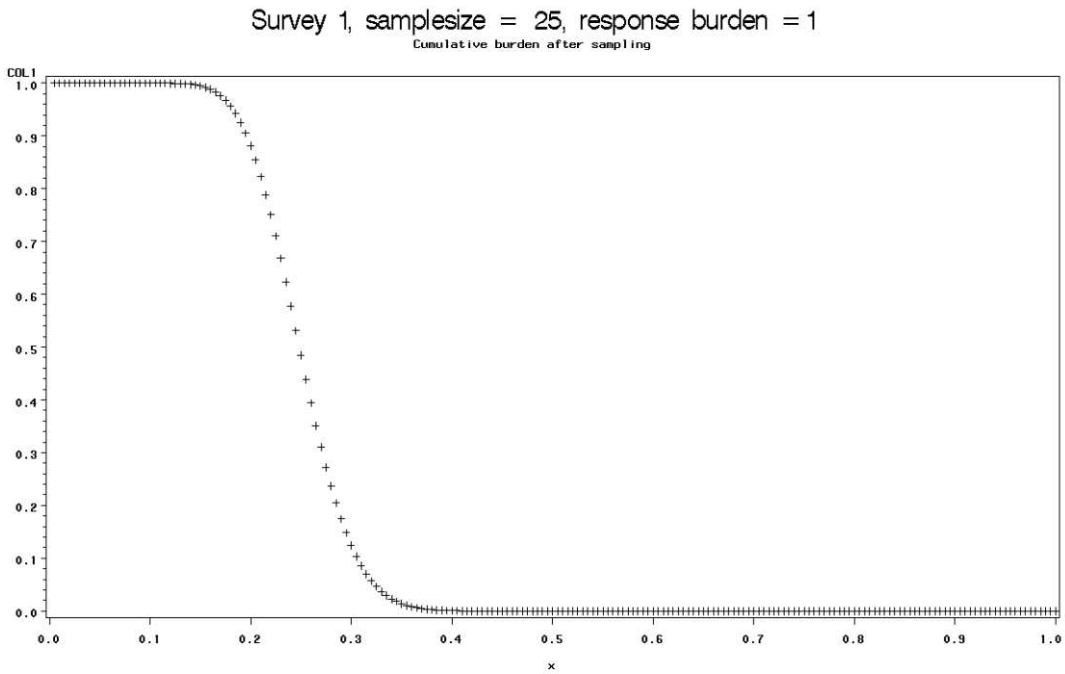
Samples n°3 and n°15 : sample size = 50, response burden = 3

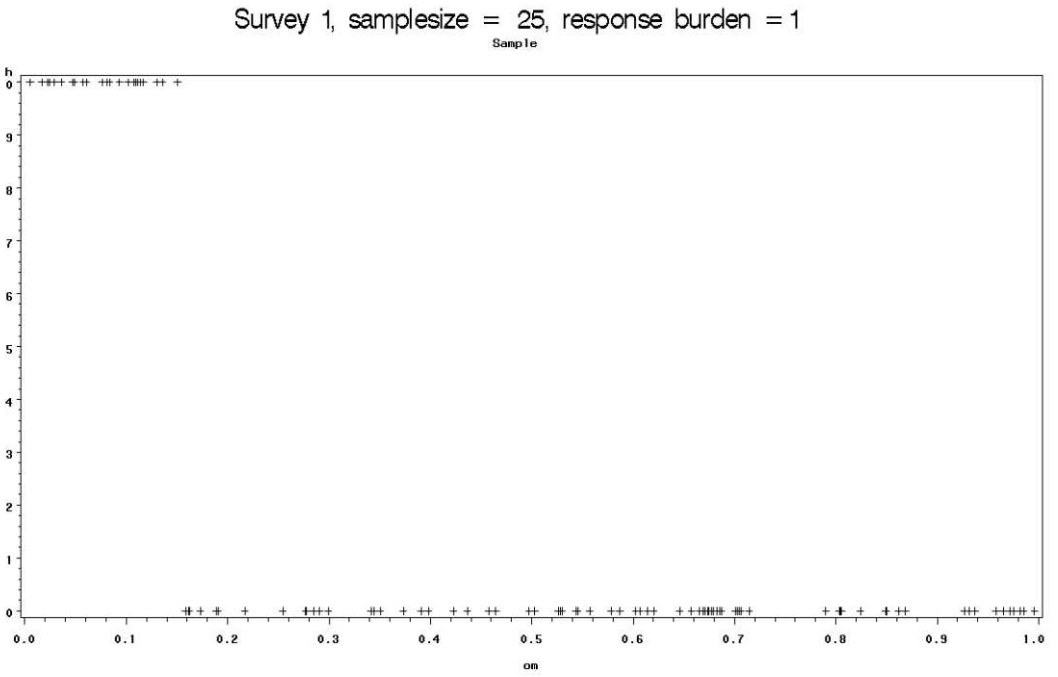
Samples n°10 and n°11 : sample size =10, response burden = 2

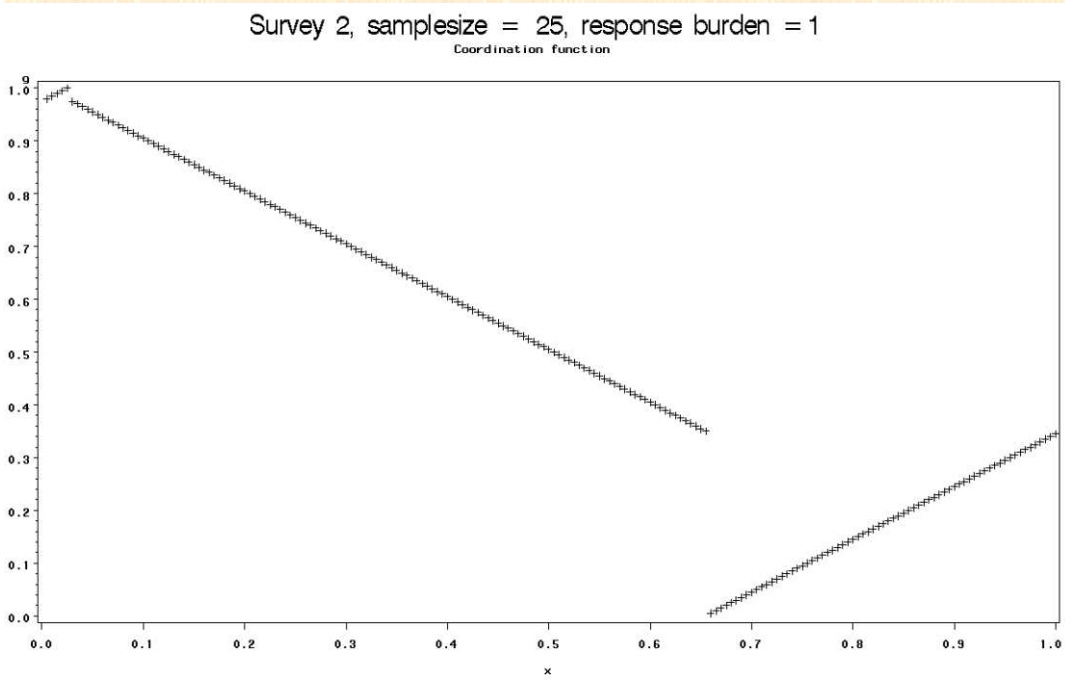
→mean cumulative burden = 7.4

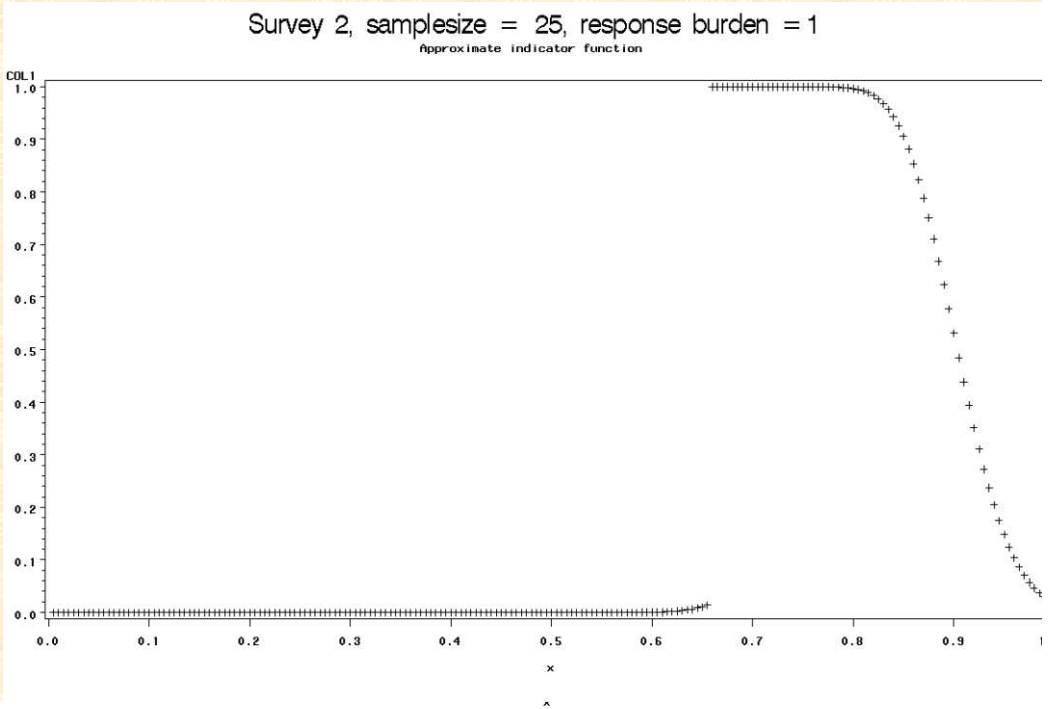

Survey 1, samplesize = 25, response burden = 1
Coordination function
x

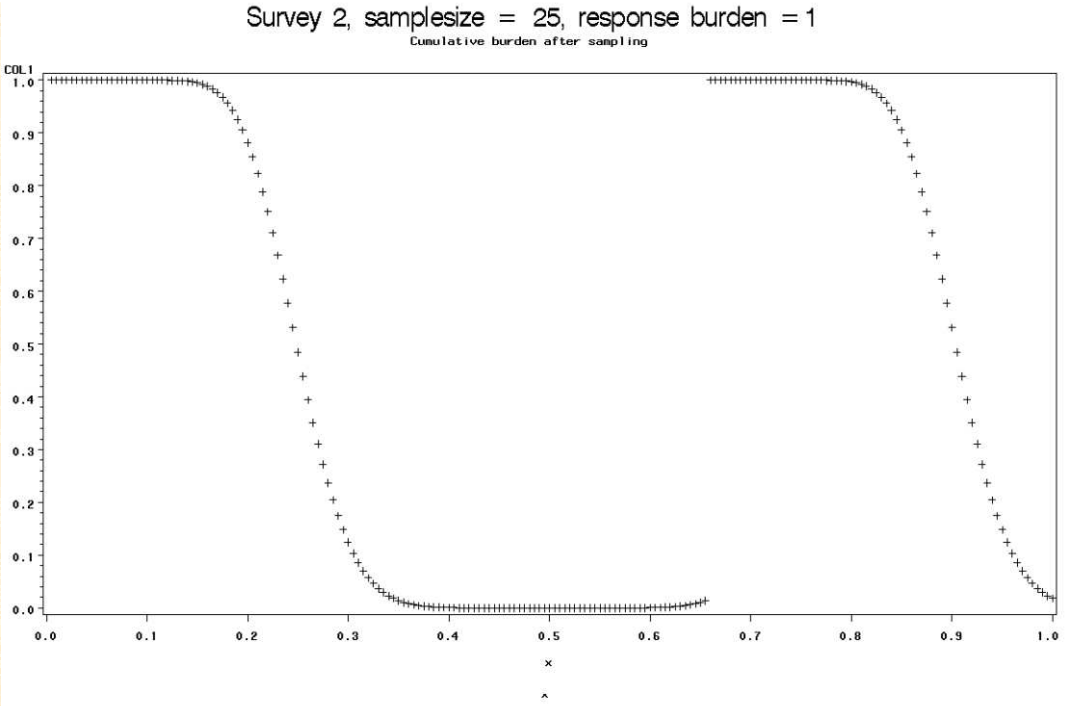

Survey 1, samplesize = 25, response burden = 1
Approximate indicator function
COL1
x

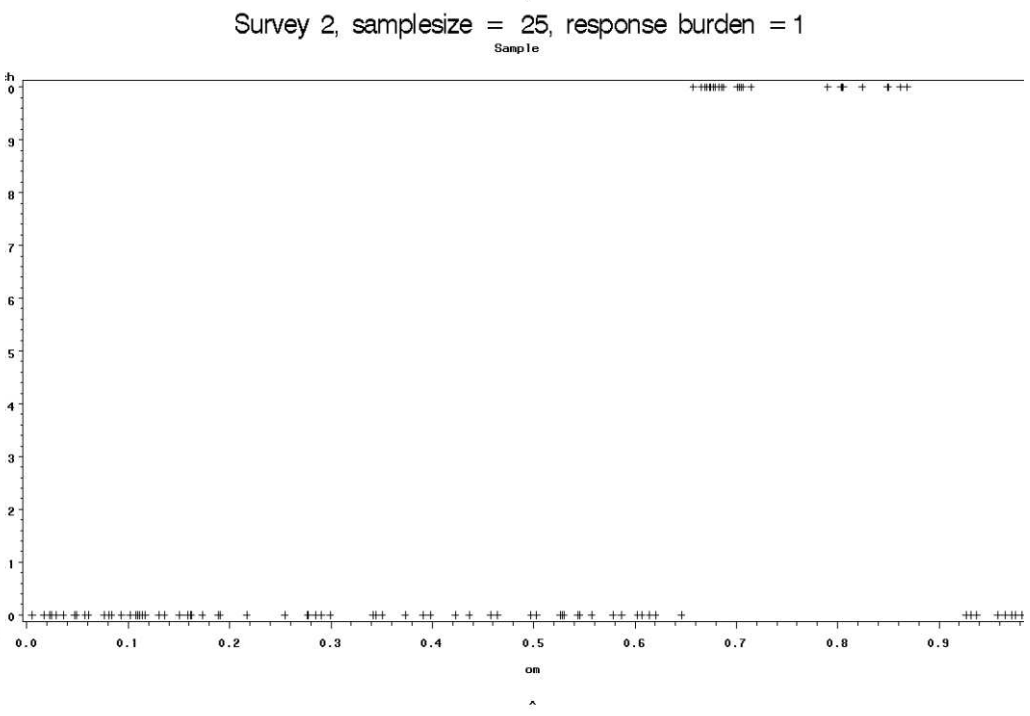

Survey 1, samplesize = 25, response burden = 1
Cumulative burden after sampling
COL1
x


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


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

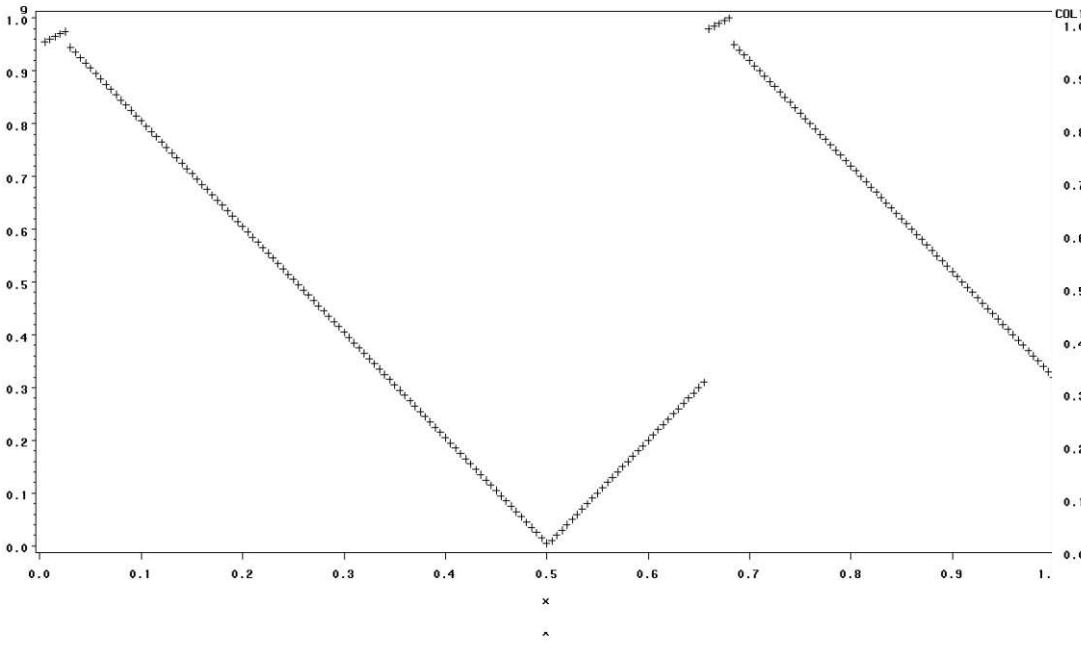

Survey 2, samplesize = 25, response burden = 1
Approximate indicator function
COL1
x


Survey 2, samplesize = 25, response burden = 1
Cumulative burden after sampling
COL1
x


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

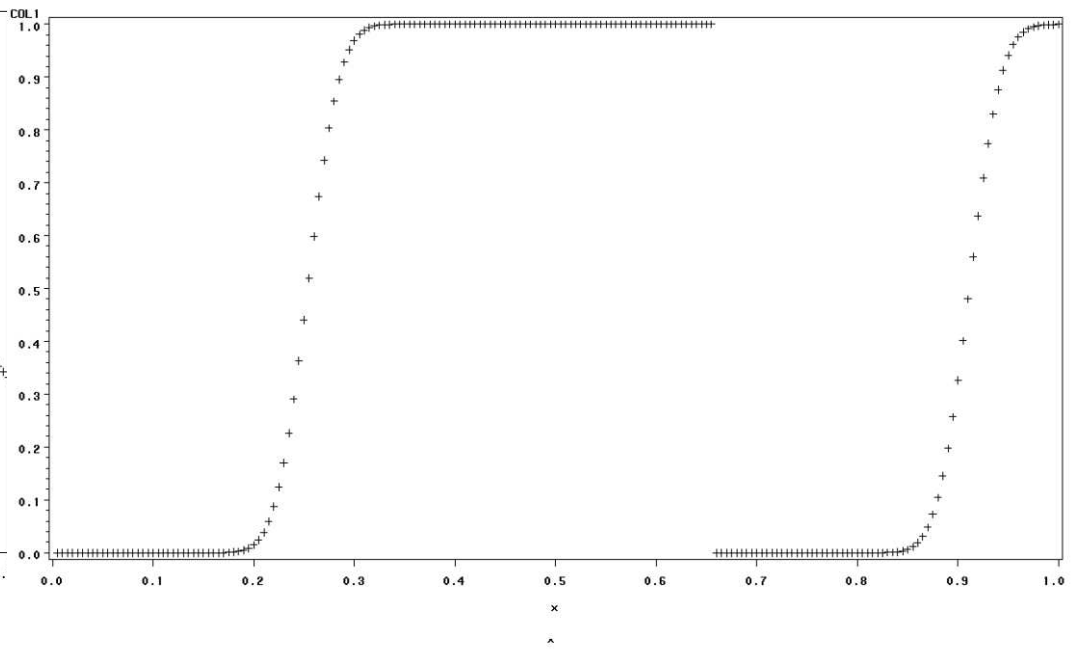
Survey 3, samplesize = 50, response burden = 3

Coordination function



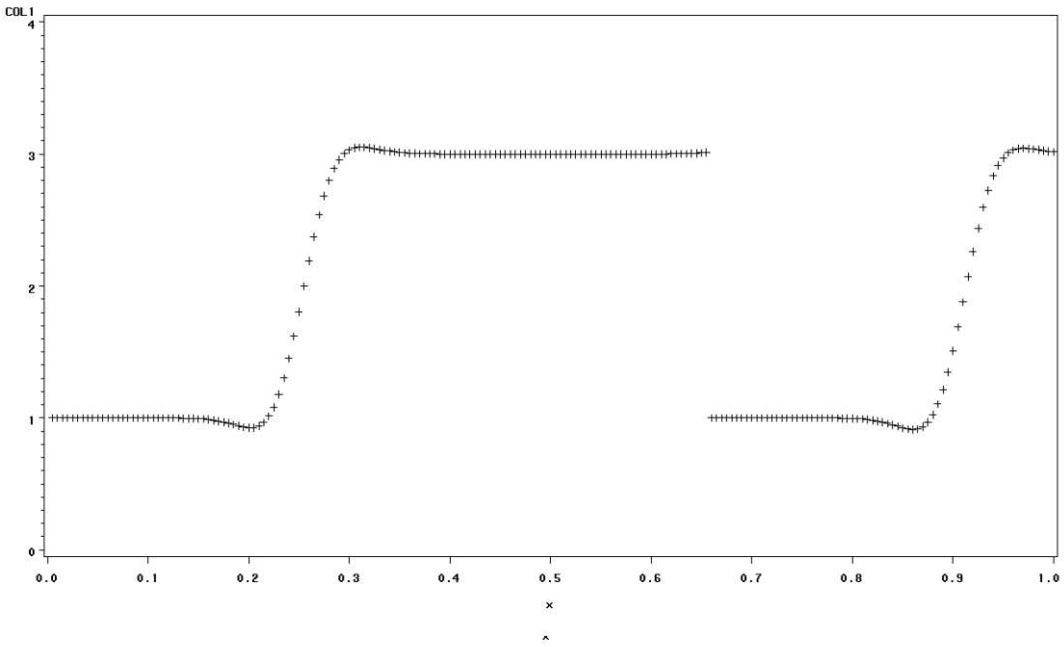
Survey 3, samplesize = 50, response burden = 3

Approximate indicator function



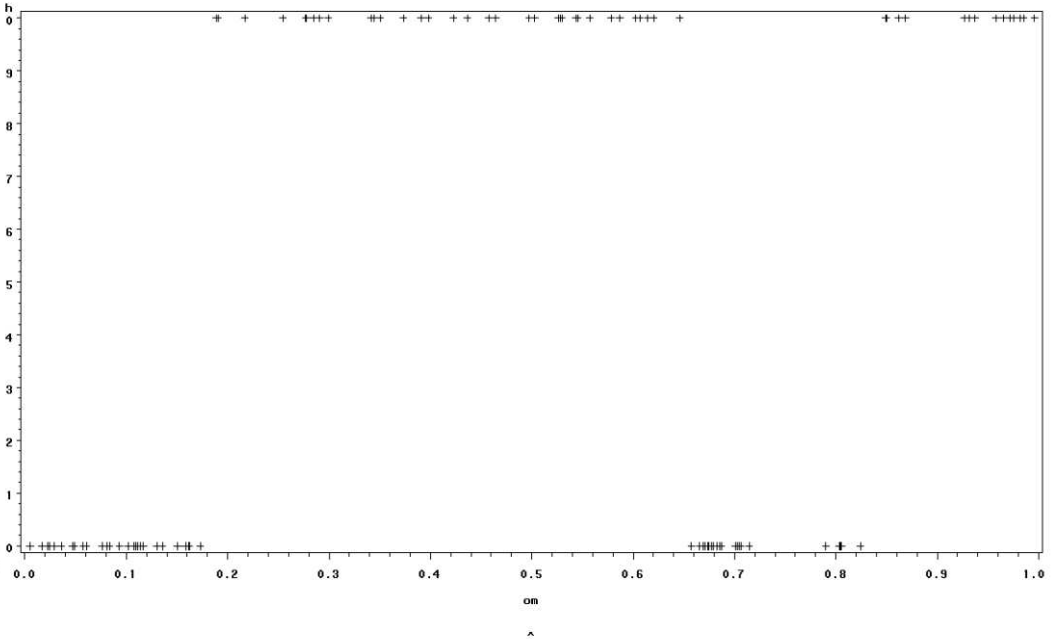
Survey 3, samplesize = 50, response burden = 3

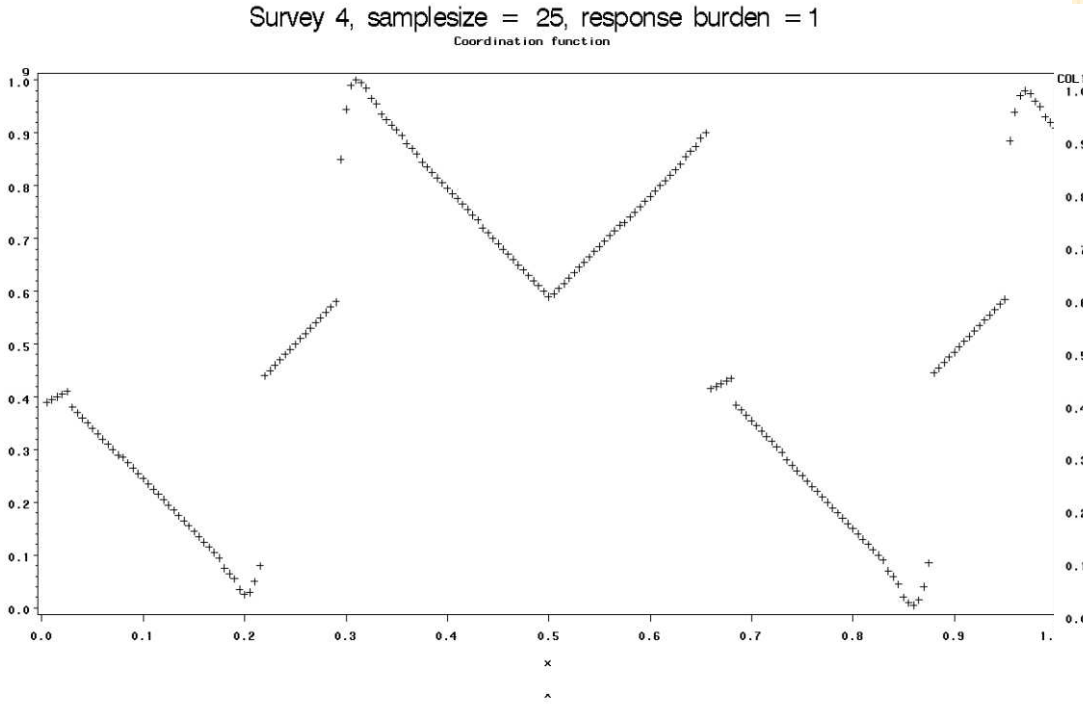
Cumulative burden after sampling

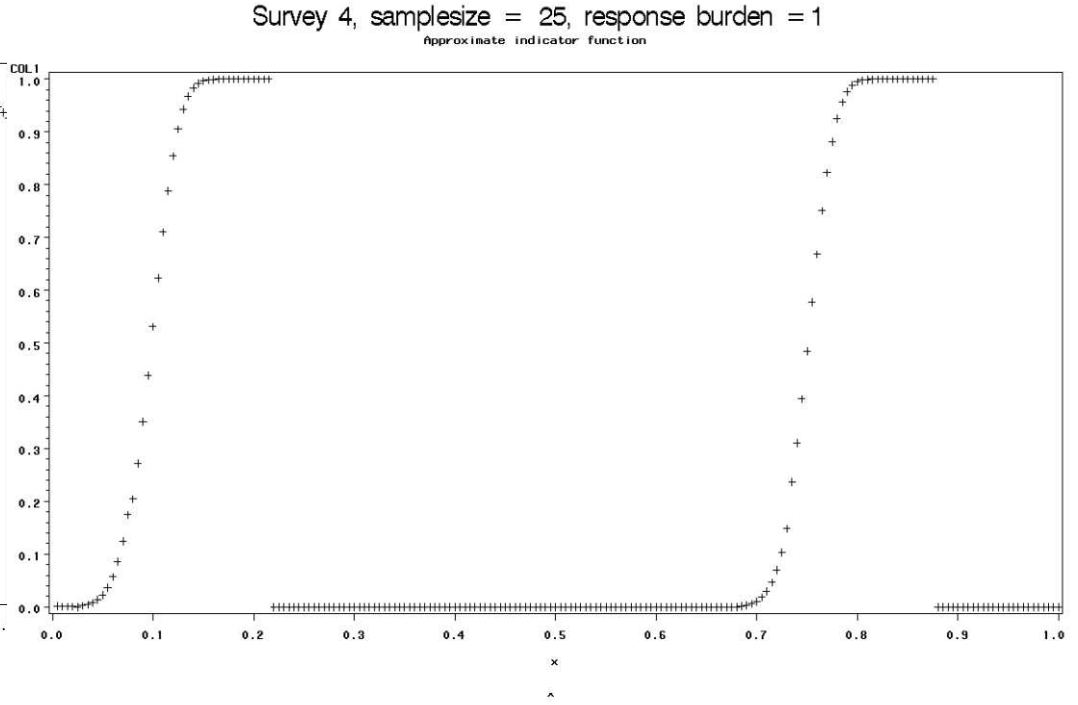


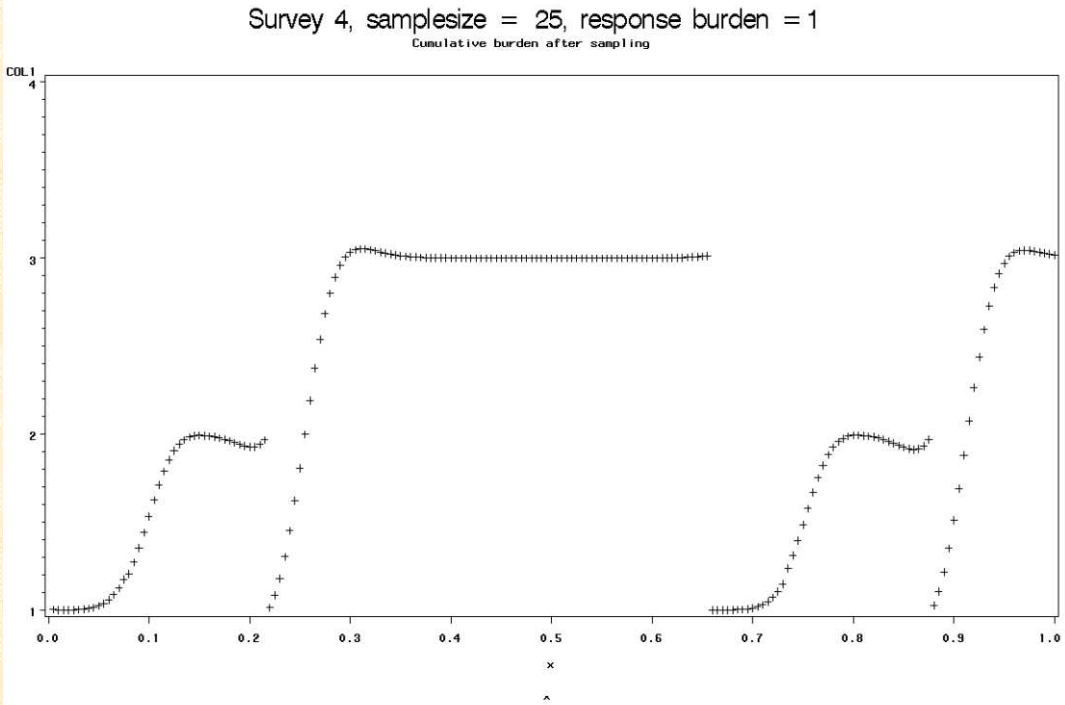
Survey 3, samplesize = 50, response burden = 3

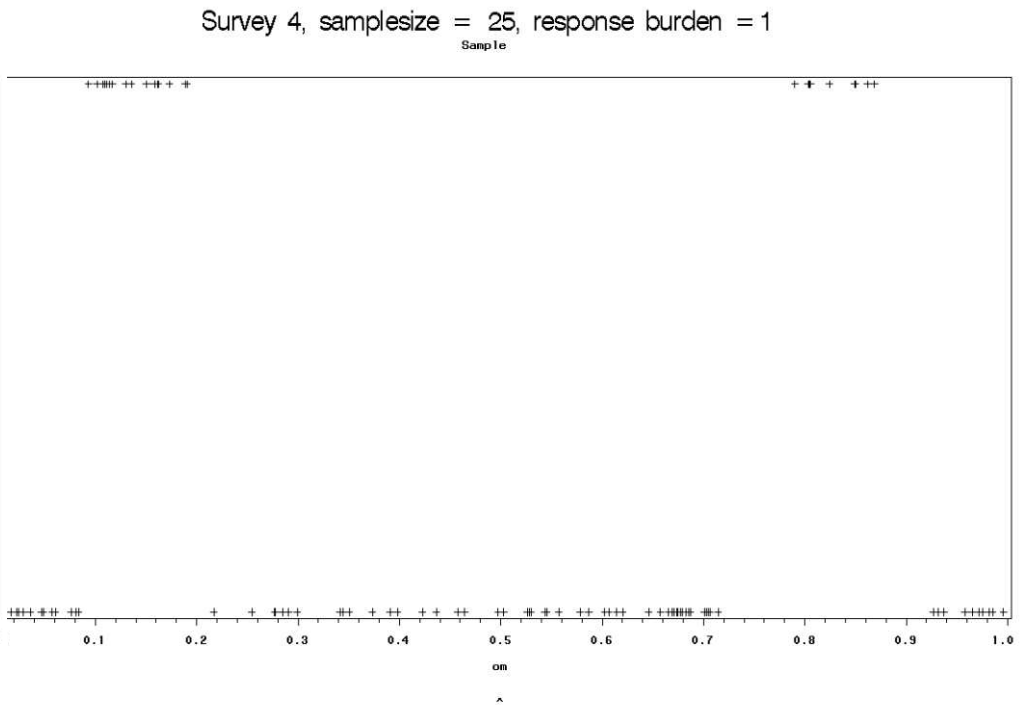
Sample

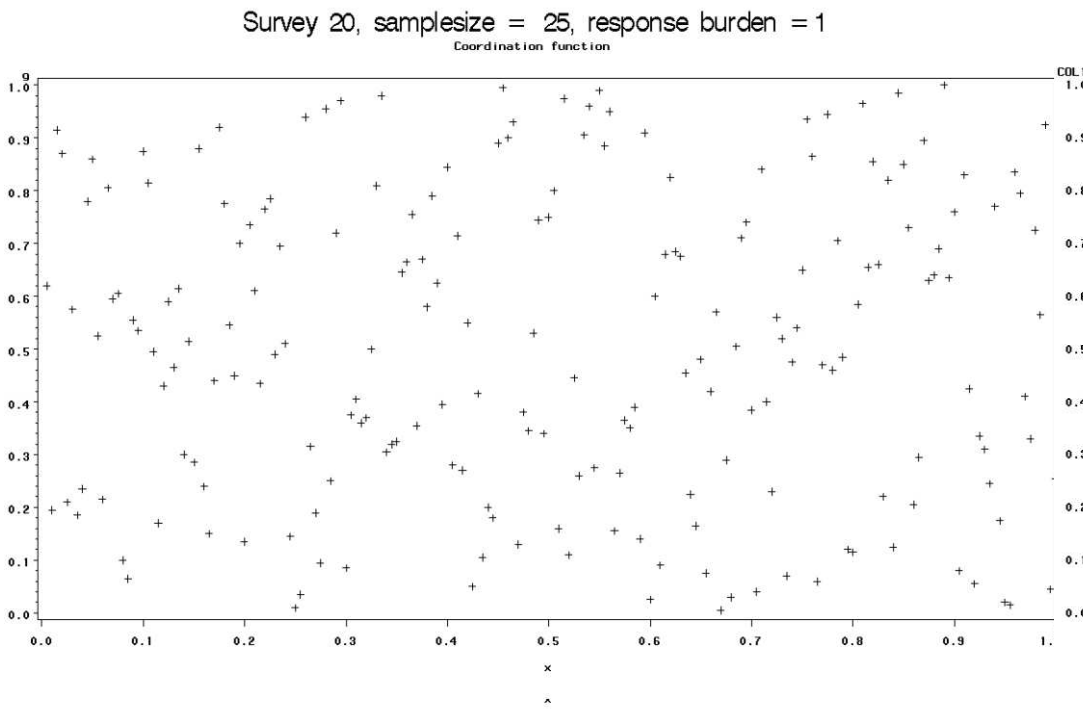


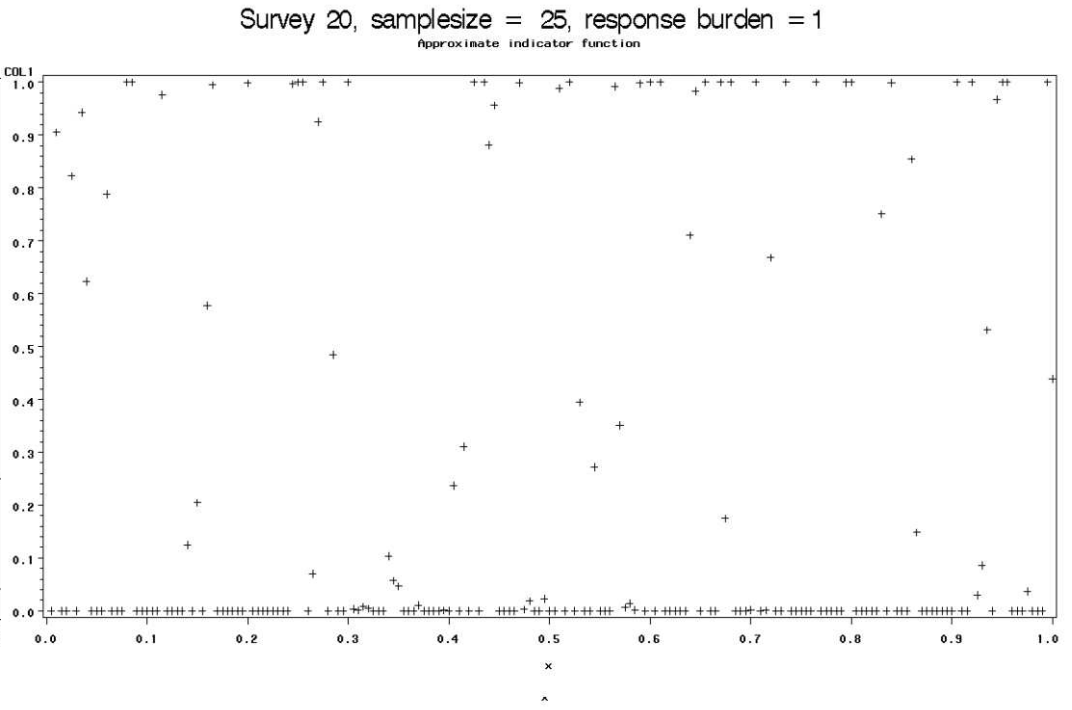

Survey 4, samplesize = 25, response burden = 1
Coordination function
x

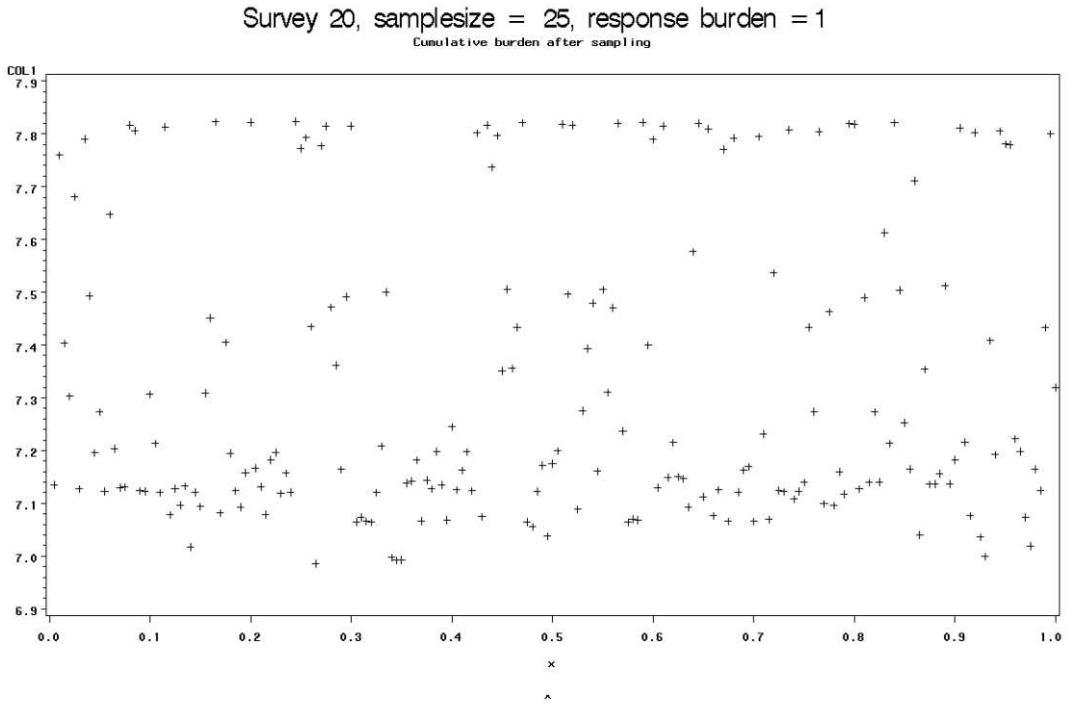

Survey 4, samplesize = 25, response burden = 1
Approximate indicator function
COL1
x

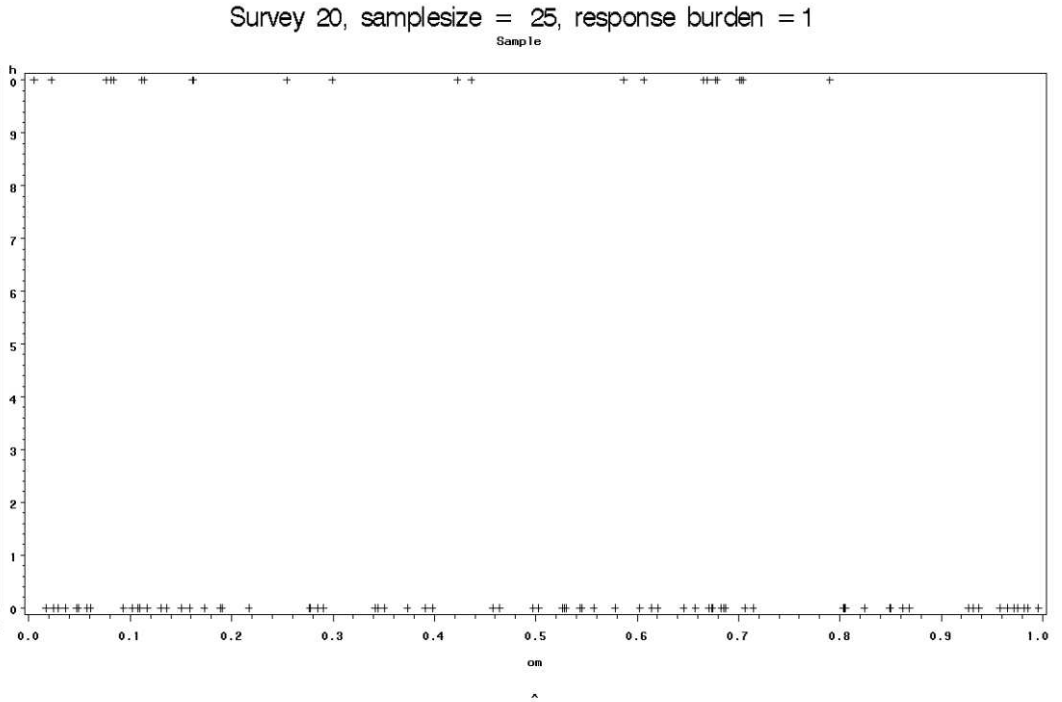

Survey 4, samplesize = 25, response burden = 1
Cumulative burden after sampling
COL1
x


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
COL1
x


Survey 20, samplesize = 25, response burden = 1
Cumulative burden after sampling
COL1
x


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