Convergence uniforme de la fonction de répartition dans le cas d'un plan de sondage informatif

D. BONNÉRY[1], F. J. BREIDT[2] and F. COQUET[1]

1. CREST-ENSAI, Campus de Ker-Lann, France
2. Colorado State University, Fort Collins, USA

Ce travail a bénéficié d'un financement de la US National Science Foundation (SES-0922142) et de l'Ensai

09/11/2010

Plan

- Quel intérêt ?
- Résultat : convergence uniforme de la fonction de répartition.
- Exemples

Un échantillon est le résultat de deux processus aléatoires qui peuvent être dépendants :

- Un **modèle de superpopulation** genère des valeurs $(Y_{.k})_{k\in[\![1,N]\!]}$ sur une population de taille N.
- Des individus sont sélectionnés via un plan de sondage aléatoire pour constituer un échantillon de taille aléatoire n.

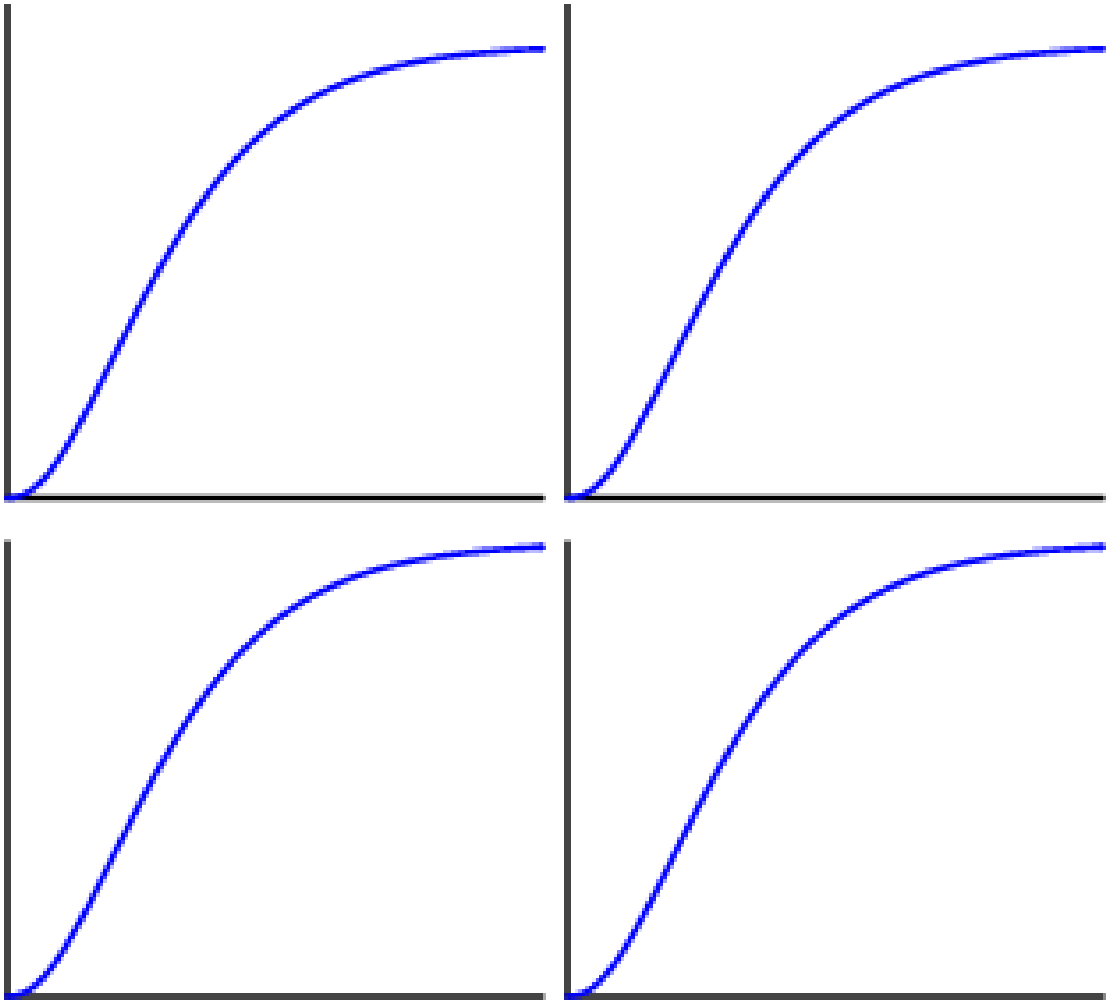
Il y a deux façons naturelles de modéliser le premier processus :

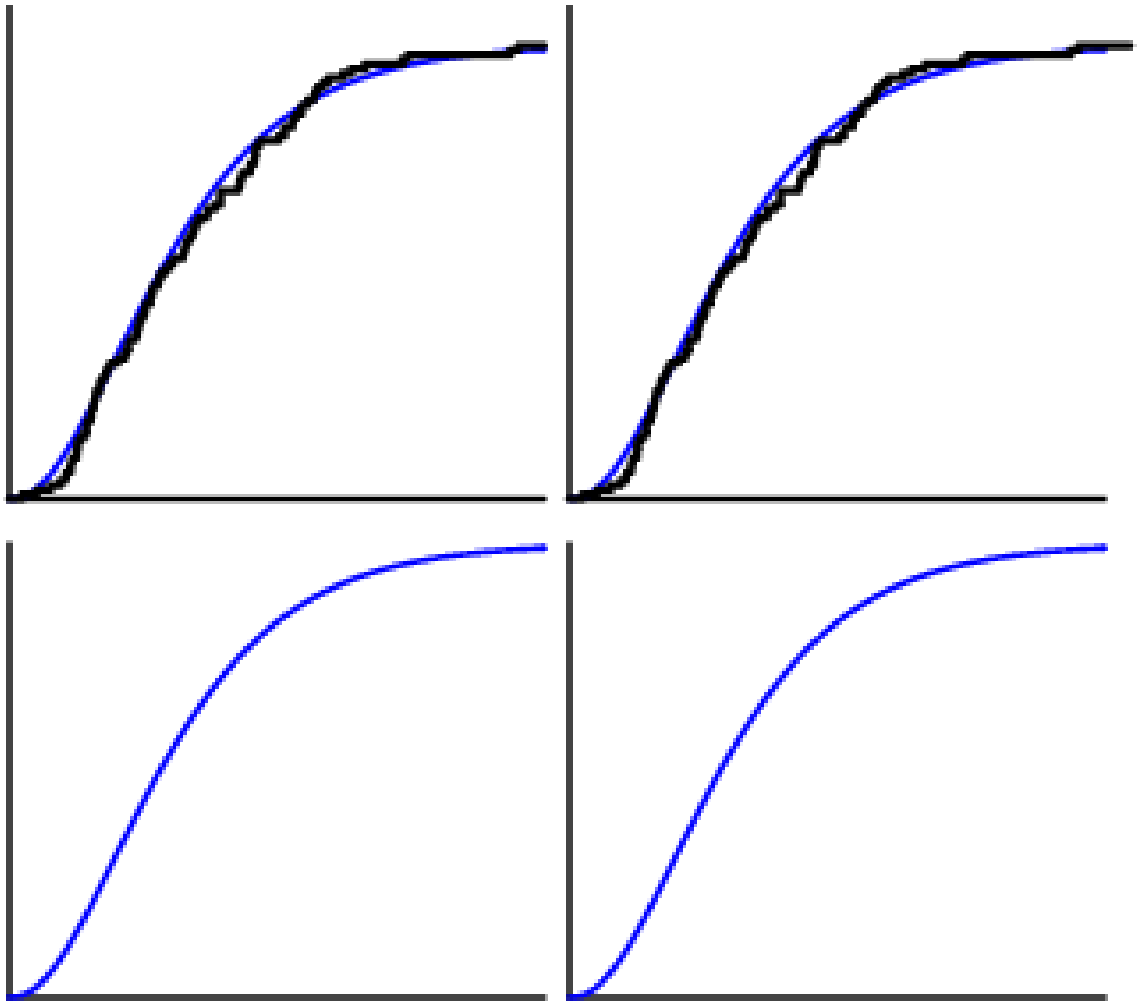
- $(Y_{.k})_{k\in[\![1,N]\!]}$ est constitué de N réalisations i.i.d. d'une même loi.(supposée appartenir à un modèle paramétrique dominé par une certaine mesure). Par la suite on supposera que $Y_{.k}$ suit une loi de densité f_θ par rapport à λ, la mesure de Lebesgue.
- $(Y_{.k})_{k\in[\![1,N]\!]}$ est le réarrangement aléatoire d'un vecteur $(y_{.k})_{k\in[\![1,N]\!]}$ qui joue le rôle de paramêtre.

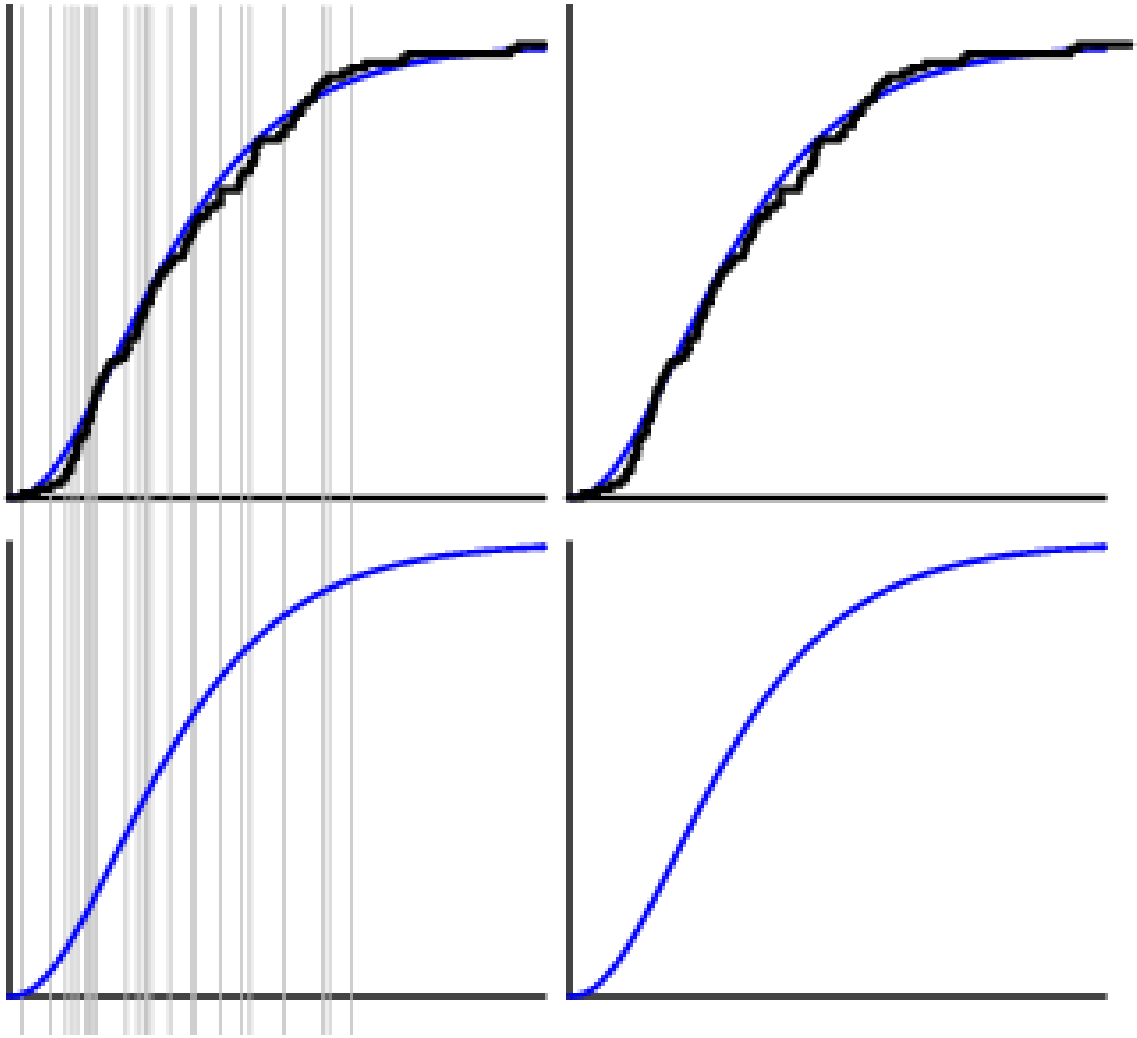
Ces deux approches se distinguent par les choix d'inférence statistique sous-jacents : sur $g(\theta)$?, $g(y_1 \dots y_N)$?.

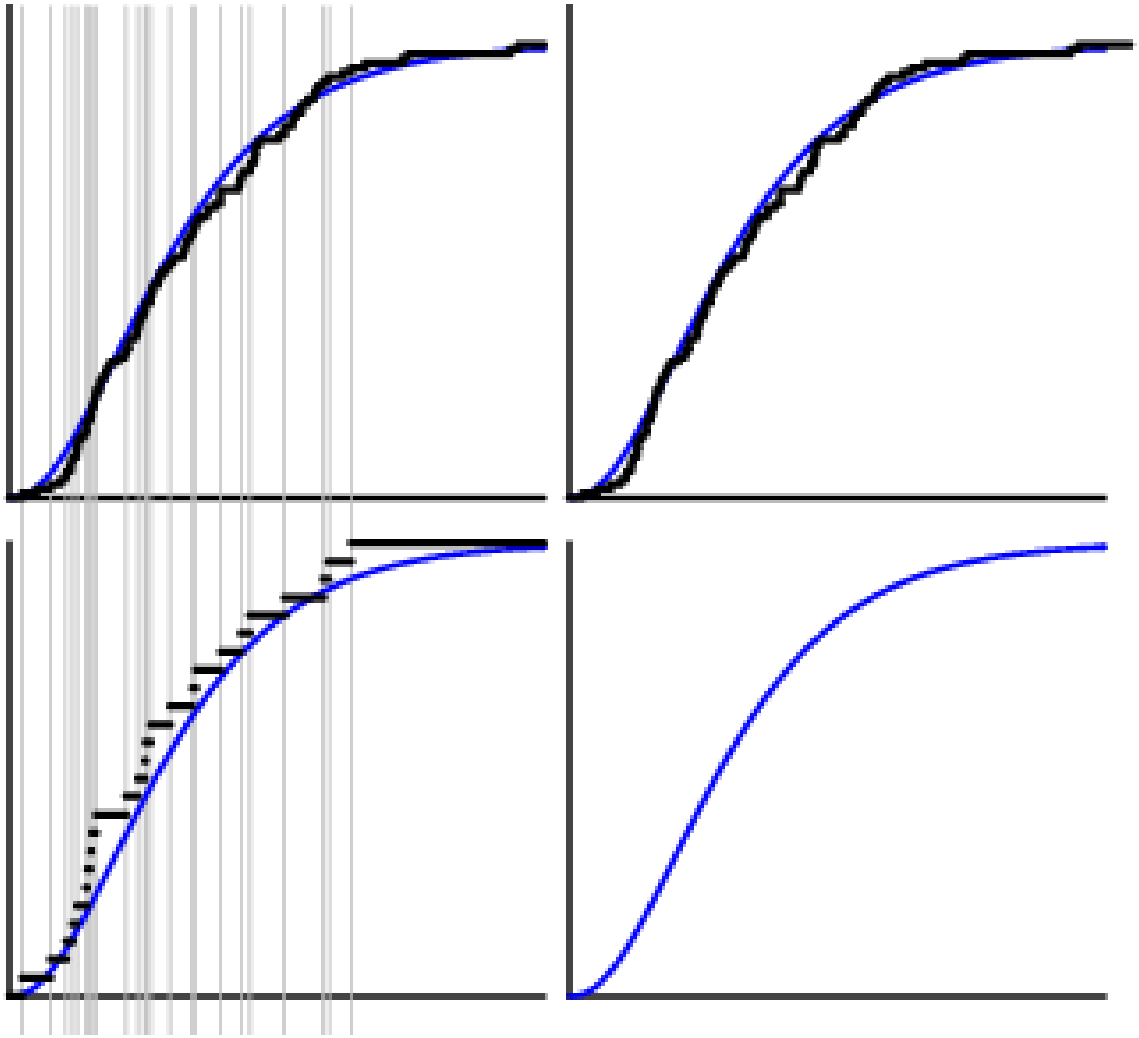
Un échantillon est un vecteur $(I_{.1}, \dots I_{.N})$ où $I_{.k}$ représente le nombre de fois que l'individu k est sélectionné.
Le plan de sondage est une variable aléatoire $\Pi : [\![1, N]\!]^{\mathbb{N}} \to [0, 1]$ qui donne la loi du vecteur $(I_{.1}, \dots I_{.N})$ conditionnellement à la réalisation de certaines variables aléatoires auxiliaires qui peuvent être liées à $(Y_{.k})_{k \in [\![1,N]\!]}$.
Le plan de sondage est **informatif** si

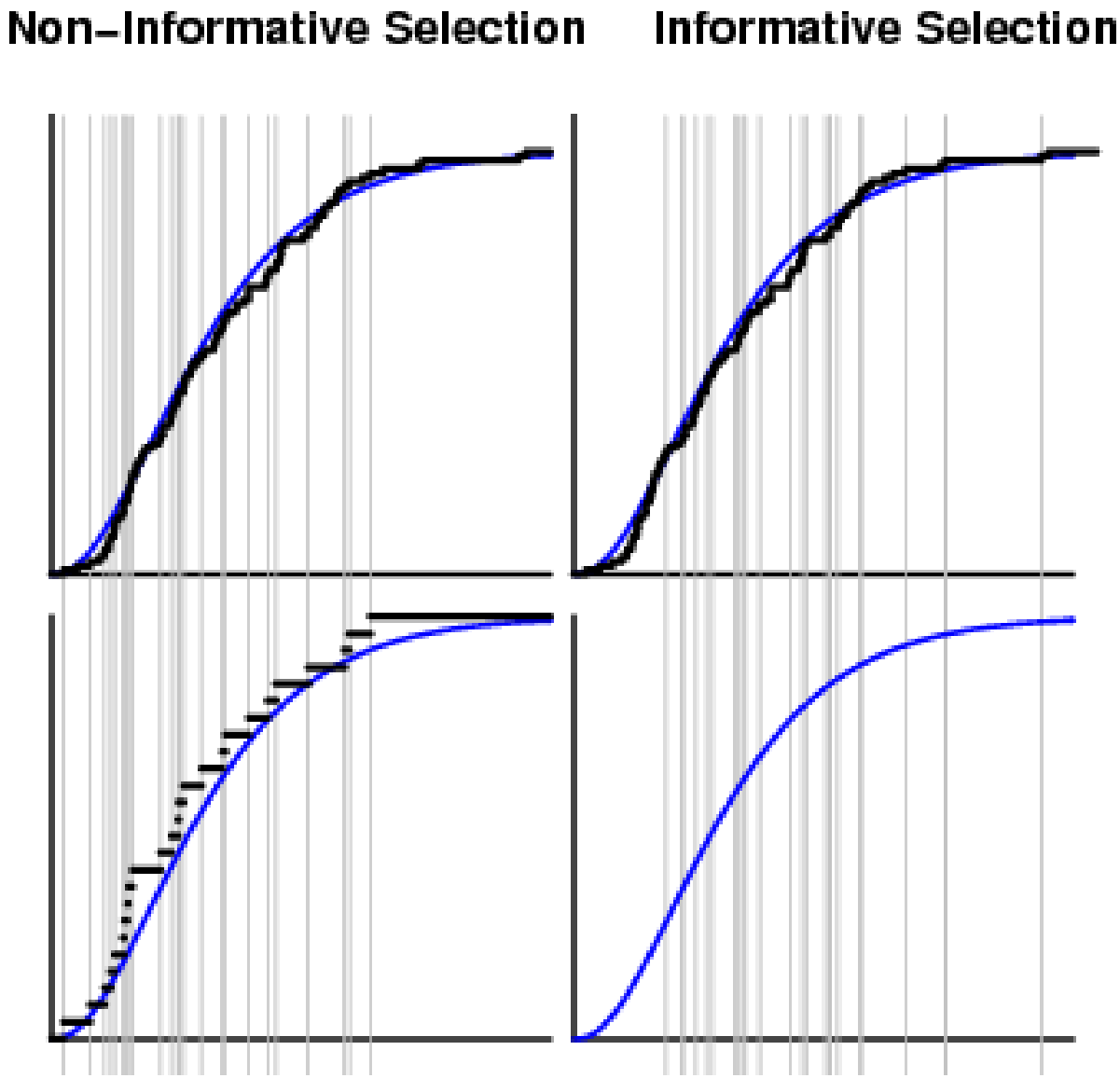
- $(I_{.1}, \dots I_{.N})$ dépend de $(Y_{.k})_{k \in [\![1,N]\!]}$
- On doit tenir compte de cette dependance lors de l'inférence sur le modèle de superpopulation.

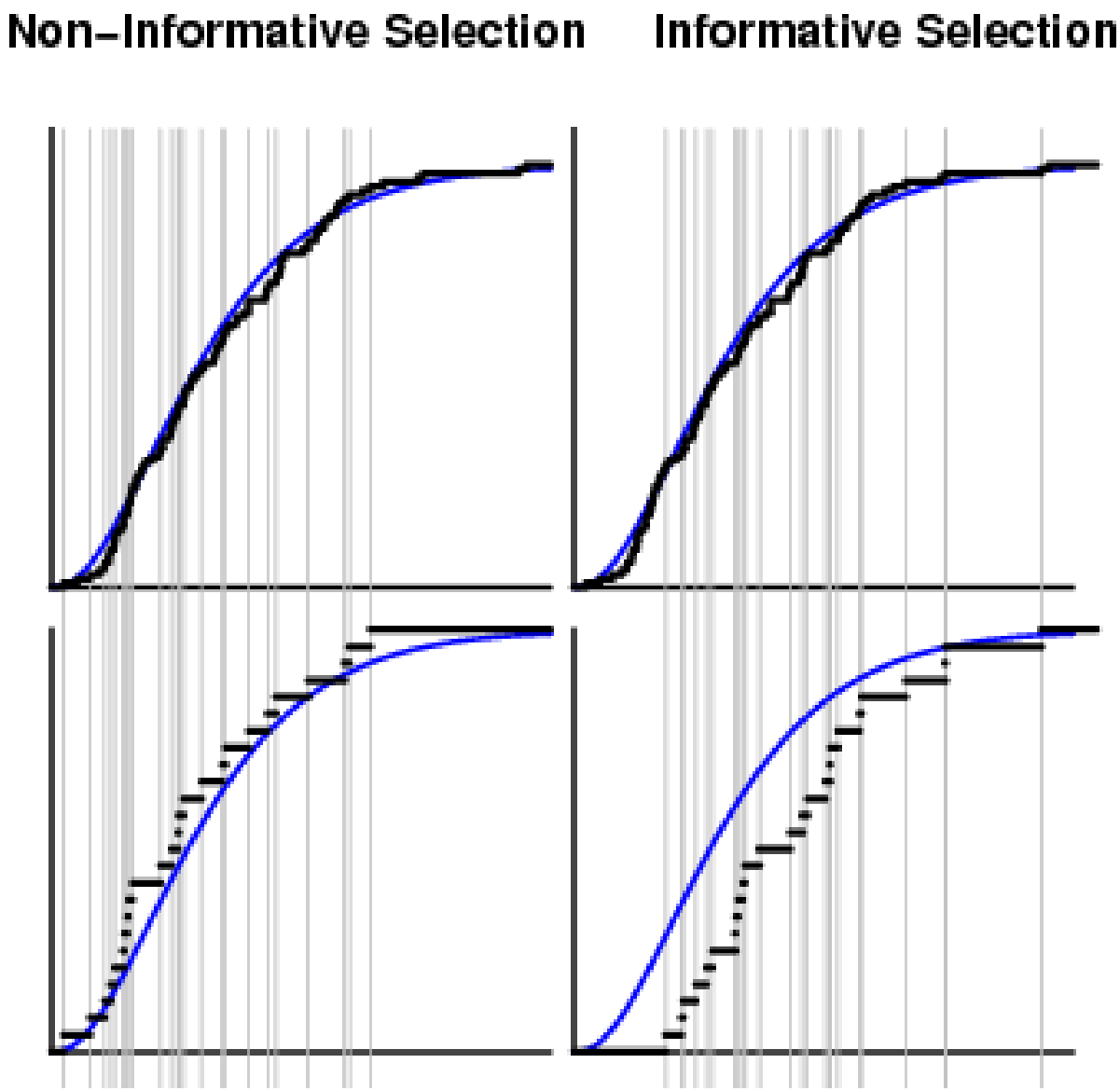


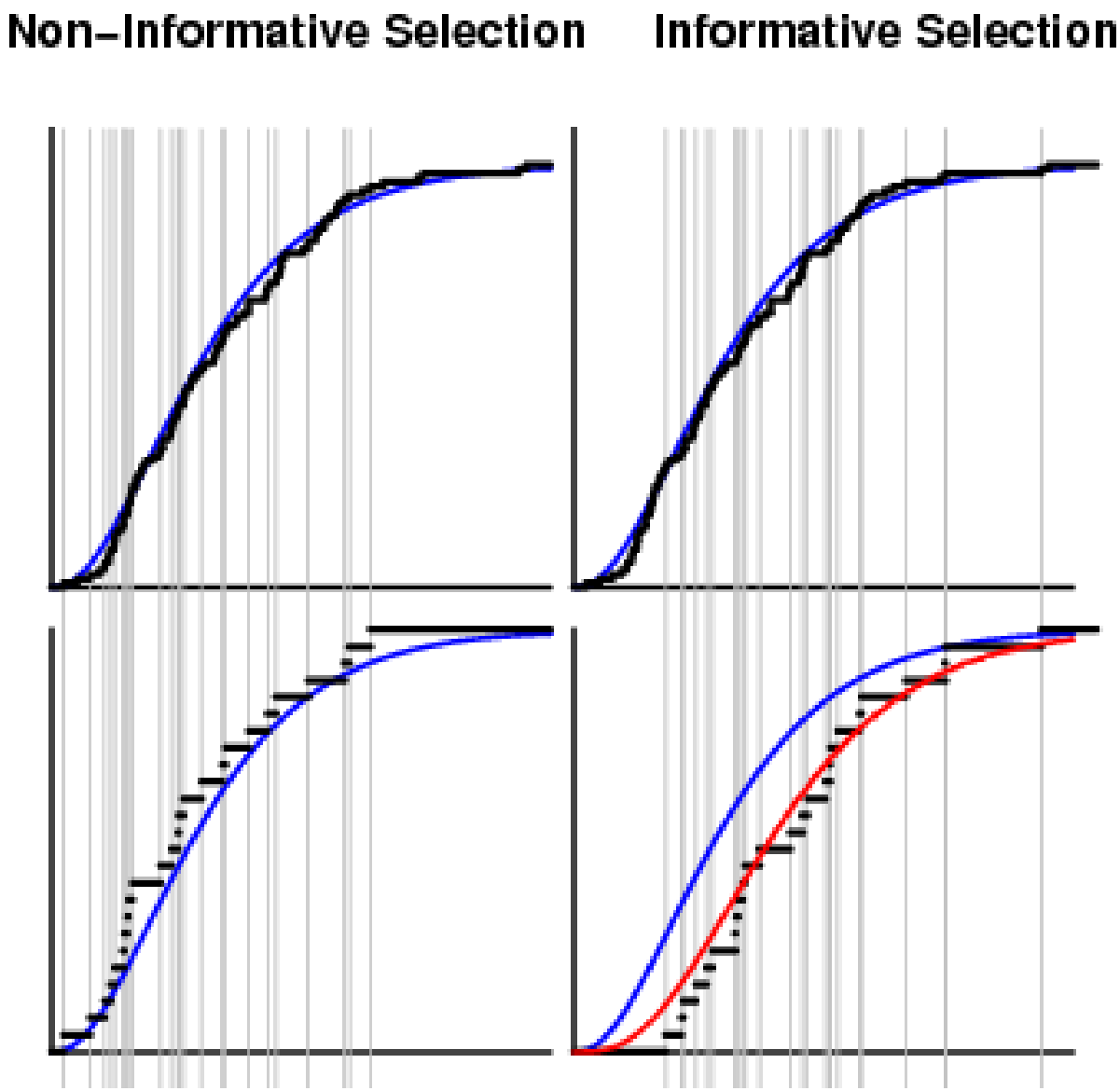







Non–Informative Selection
Informative Selection


Non–Informative Selection
Informative Selection


Non–Informative Selection
Informative Selection

- Pfeffermann et al. ont développé des méthodes d'estimation basées sur une approximation de la vraisemblance
  - comme produit de densitées pondérées
  - sous de fortes hypothèses, et un cadre restrictif (n fixé, $N \to \infty$)

  Nous voulons prendre du recul, pour commencer à répondre aux questions :
- Dans le cas d'un sondage informatif, la vraisemblance approchée est elle valable ?
- Quelles sont les propriétés des statistiques de bases liées à cette densité ?

$(N_\gamma)_{\gamma\in\mathbb{N}}$ est une suite croissante d'entiers, $(y_k)_{k\in\mathbb{N}}$ une suite de réels. Les variables aléatoires sont définies sur l'espace probabilisé $(\Omega, \mathscr{A}, P)$.

$$\begin{array}{c|ccccccccc} Y & Y_{.1} & \dots & Y_{.N_1} & \dots & Y_{.N_2} & \dots & Y_{.N_\gamma} & \dots \\ I_1 & I_{11} & \dots & I_{1N_1} & & & & & \\ I_2 & I_{21} & \dots & \dots & \dots & I_{2N_2} & & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ I_\gamma & I_{\gamma 1} & \dots & I_{\gamma k} & \dots & \dots & \dots & I_{\gamma N_\gamma} & \end{array}$$

Pour $\gamma \in \mathbb{N}$,

- un échantillon est un vecteur $I_\gamma \in \mathbb{N}^{N_\gamma}$, $I_{\gamma k}$ est le nombre de fois que k est sélectionné.
- $Y_\gamma = (Y_{.k})_{k\in[\![1,N_\gamma]\!]}$

On suppose l'échangeabilité par colonne : $\forall\gamma \in \mathbb{N}$, σ une permutation de $[\![1, N_\gamma]\!]$, (I_γ, Y_γ) et $(\sigma.I_\gamma, \sigma.Y_\gamma)$ sont identiquement distribués.

Definition

Soit $\gamma \in \mathbb{N}$, la fonction de répartition empirique de l'échantillon est le processus aléatoire

$$\begin{array}{rccl} F_\gamma : \omega \mapsto F_\gamma(\omega) & : & \mathbb{R} & \to & [0,1] \\ & & \alpha & \mapsto & \frac{\sum_{k=1}^{N_\gamma} \mathbb{1}_{]-\infty,\alpha]}(Y_k) I_{\gamma k}}{n_\gamma + \mathbb{1}_{I_\gamma = 0}} \end{array}$$

On définit $m_\gamma : y \mapsto E[I_{\gamma k} | Y_{.k} = y]$
La densité d'une observation sélectionnée est la fonction : $w_\gamma f$, où $w_\gamma(y) = \frac{m_\gamma(y)}{\int m_\gamma f d\lambda}$ et la fonction de répartion des observations sélectionnées est $y \mapsto \int_{]-\infty,\alpha]} w_\gamma f d\lambda$

Propriété

$w_\gamma f$ est une densité de probabilité.

On note $Y_{.(k)}$ le $k$ème individu tiré. Le vecteur des observations est $(Y_{.(k)})_{k\in[\![1,n_\gamma]\!]}$.

Propriété

- *Dans le cas d'un tirage à taille n_γ fixe,*

$$\forall k, P^{Y_{.(k)}} = w_\gamma f.\lambda$$

et $E\left[F_\gamma(\alpha)\right] = \int_{]-\infty,\alpha]} w_\gamma f d\lambda$

- *Dans le cas d'un tirage sans remise :*

$$P^{Y_{.k}|I_{\gamma k}=1} = w_\gamma f.\lambda$$

Résultats : convergence uniforme de la fonction de répartition empirique

$$m_\gamma(y) = E[I_{\gamma k}|Y_{.k} = y]$$

Hypothèses

$$A_0 \begin{cases} \exists\ M \in L_1(\lambda) s.t.\ \forall \gamma, m_\gamma < M \\ \forall y \in \mathbb{R}, m(y) = \lim_\gamma m_\gamma(y) \textit{exists} \\ \int mf\, d\lambda > 0 \end{cases}$$

Definition

Pour $\gamma \in \mathbb{N}$, $\alpha \in \mathbb{R}$, on définit la densité et fonction de répartition limites lorsque A_0 est vérifié, comme

$$\begin{aligned} w(\alpha) &= \lim_\gamma w_\gamma = \lim_\gamma \frac{m_\gamma(\alpha)}{\int m_\gamma f d\lambda} = \frac{m(\alpha)}{\int mf d\lambda} \\ F_s(\alpha) &= \frac{\int \mathbb{1}_{(-\infty,\alpha]} mf\, d\lambda}{\int mf\, d\lambda} = \int_{-\infty}^{\alpha} wf\, d\lambda \end{aligned}$$

On définit:

$$
\begin{aligned}
&m_\gamma(y) = E[I_{\gamma 1}|Y_{.1} = y] \\
&m'_\gamma(y_1, y_2) = E[I_{\gamma 2}|Y_{.1} = y_1, Y_{.2} = y_2] \\
&v_\gamma(y) = V[I_{\gamma 1}|Y_{.1} = y] \\
&c_\gamma(y_1, y_2) = Cov[I_{\gamma 1}, I_{\gamma 2}|Y_{.k_1} = y_1, Y_{.2} = y_2]
\end{aligned}
$$

Hypothèses

$$
A_1 \begin{cases}
\int c_\gamma(y_1, y_2) f(y_1) f(y_2) dy_1 dy_2 = o_{\gamma\to\infty}(1) \\
\int \left(m'_\gamma(y_1, y_2) m'_\gamma(y_2, y_1) - m_\gamma(y_1) m_\gamma(y_2)\right) f(y_1) f(y_2) dy_1 dy_2 \\
\qquad = o_{\gamma\to\infty}(1) \\
\int v_\gamma \, f \, d\lambda = o_{\gamma\to\infty}(N_\gamma), \quad \int m_\gamma^2 \, f \, d\lambda = o_{\gamma\to\infty}(N_\gamma) \\
P(\{I_\gamma = (0 \dots 0)\}) = o_\gamma(1)
\end{cases}
$$

On définit la version presque sûre des hypothèses précédentes :

Hypothèses

$$A_2 : \forall y \in \mathbb{R}^{\mathbb{N}} t.q. \ \sup_\alpha \left\| \frac{\sum_{k=1}^{N_\gamma} b_\alpha(y_k)}{N_\gamma} - \int b_\alpha f \, d\lambda \right\| = o_\gamma(1), \forall \alpha \in \mathbb{R},$$

$$\begin{cases} \sum_{k,l=1}^{N_\gamma} b_\alpha(y_k) b_\alpha(y_l) Cov\left[I_{\gamma k}, I_{\gamma l} | Y_\gamma = (y_1 \dots y_{N_\gamma})\right] = o_\gamma(N_\gamma^2) \\ \sum_{k=1}^{N_\gamma} b_\alpha(y_k) \left(E\left[I_{\gamma k} | Y_\gamma = (y_1 \dots y_{N_\gamma})\right] - m_\gamma(y_k)\right) = o_\gamma(N_\gamma) \\ g_\gamma(y, (0 \dots 0)) = o_\gamma(1) \end{cases}$$

Théorème

- *Si A_0 et A_1 sont vérifiées,*
  alors
  $$\|F_\gamma - F_s\|_\infty \xrightarrow[\gamma\to\infty]{L_2} 0$$
  avec $\|F_\gamma - F_s\|_\infty = \sup_{\alpha\in\mathbb{R}} \{|F_\gamma(\alpha) - F_s(\alpha)|\}$
- *Si $A0$ et $A2$ sont vérifiées, alors il existe une suite de variables aléatoires $(I'_{\gamma k})_{\gamma\in\mathbb{N},k\in U_\gamma}$, $(Y'_k)_{k\in\mathbb{N}}$ definies sur $\left(\Omega\times[0,1], \mathscr{A}\otimes\mathscr{B}_{[0,1]}, P' = P\otimes\lambda_{[0,1]}\right)$ telles que*
  - $\|F'_\gamma - F_s\|_\infty$ *converge P'-p.s. vers* 0
  - $\forall\gamma\in\mathbb{N}, (I'_\gamma, Y'_\gamma)$ *et* (I_γ, Y_γ) *ont la même loi*
  - $\forall\gamma\in\mathbb{N}, \omega\in\Omega, x\in[0,1],\ Y'_\gamma(\omega,x) = Y_\gamma(\omega)$

  avec
  $$F'_\gamma : \mathbb{R}\to[0,1], \alpha\mapsto F'_\gamma(\alpha) = \frac{\sum_{k\in U_\gamma} \mathbb{1}_{(-\infty,\alpha]}(Y'_{\gamma k}) I'_{\gamma k}}{\sum_{k\in U_\gamma} I'_{\gamma k} + \mathbb{1}_{I'_\gamma=0}}.$$

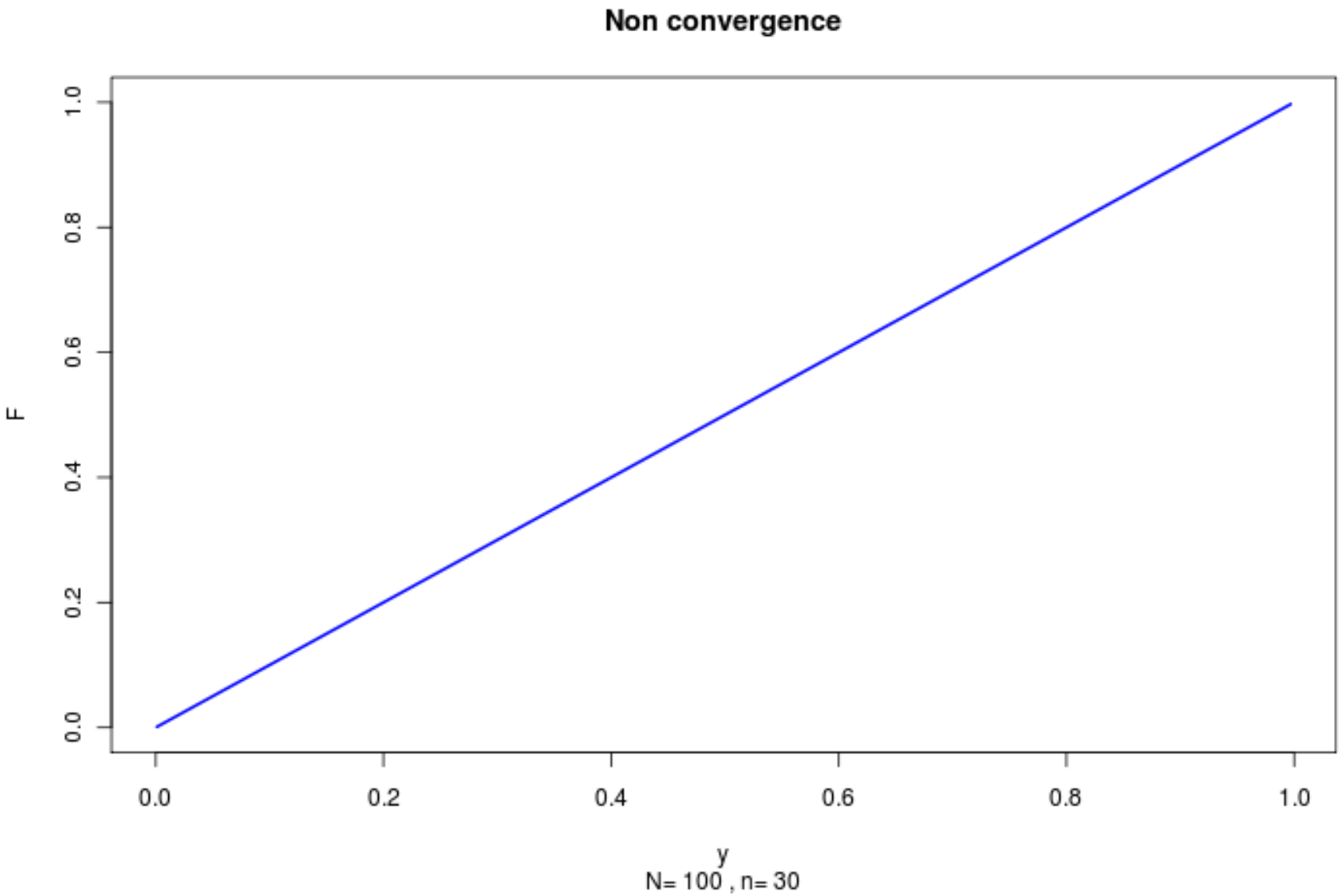
Exemples et simulations

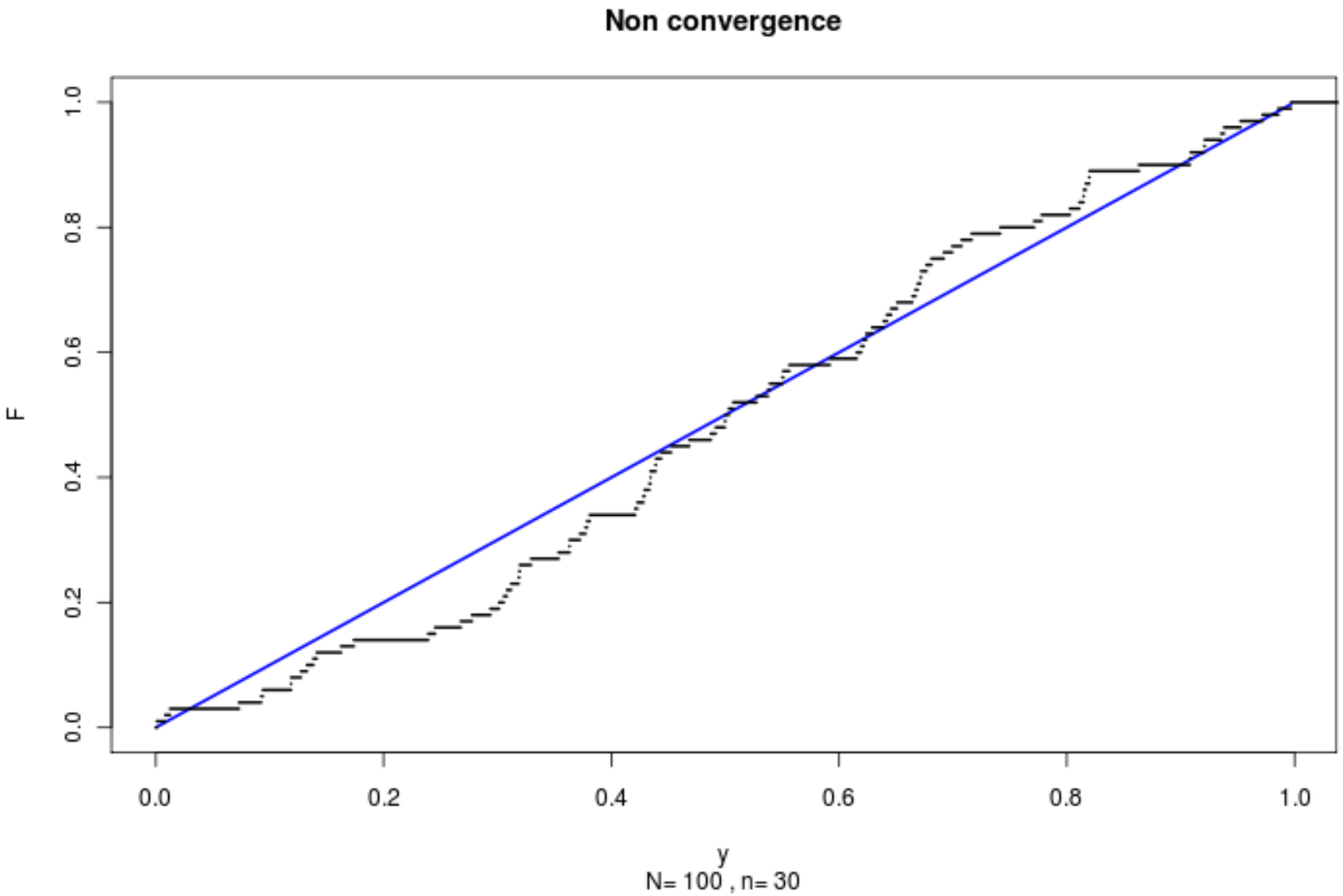
Exemple de non convergence

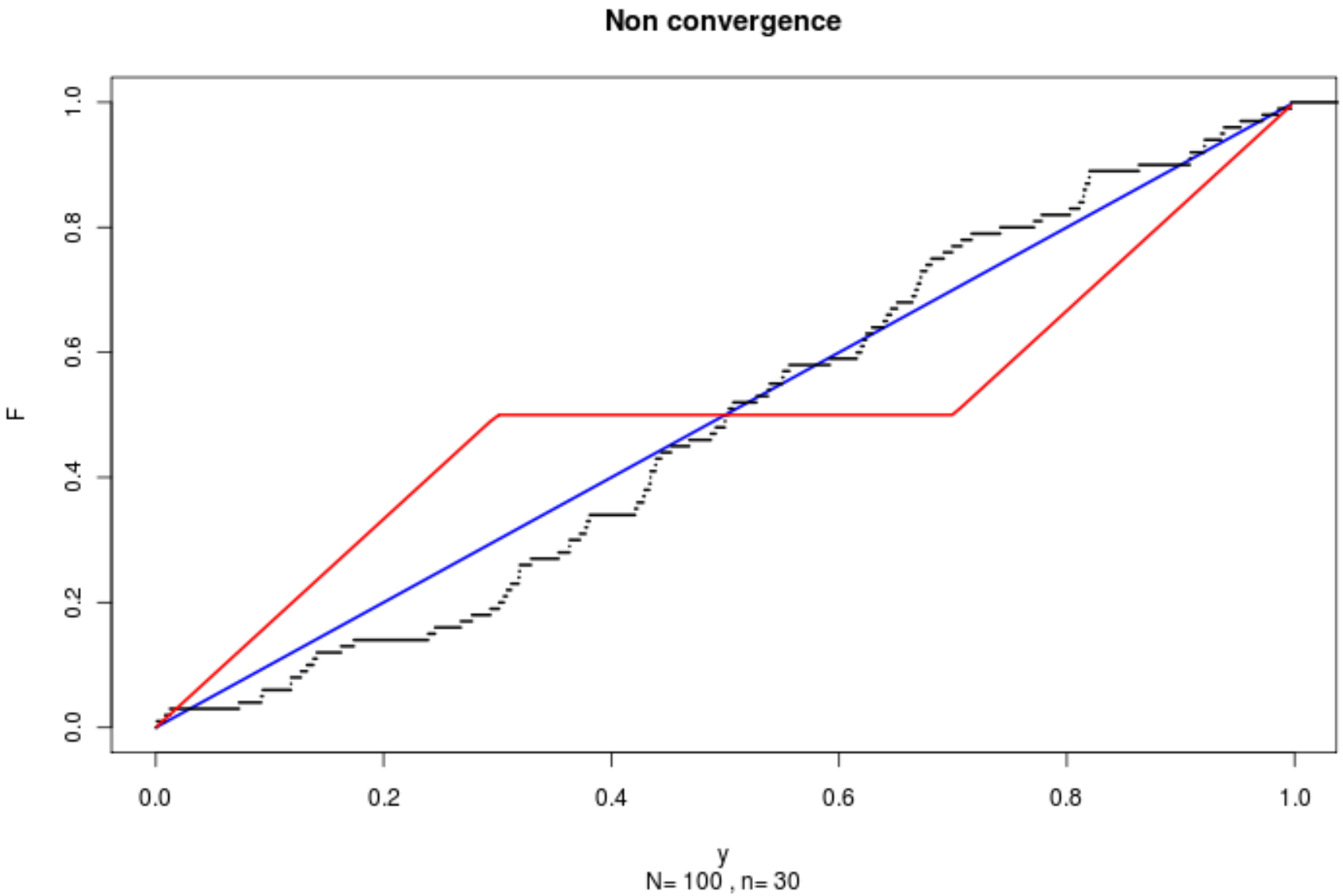
La limite F_s peut exister (A_0 est vérifiée), sans que F_γ ne converge (A_1, A_2 ne sont pas vérifiées):
On suppose que $Y \sim U_{[0,1]}$. On considère le plan de sondage suivant :

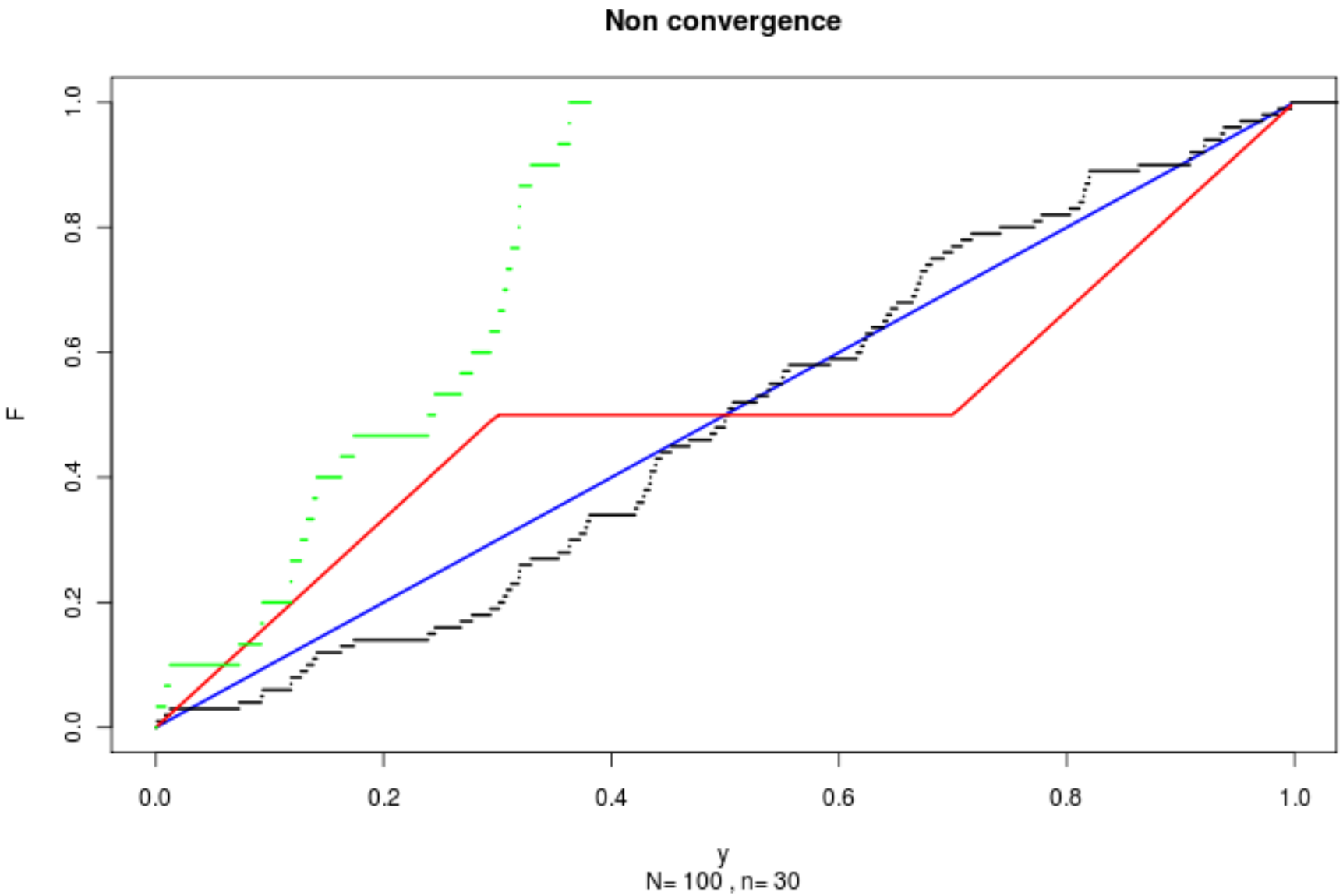
$P(I_\gamma)$	$Y_{(1)}$	$\dots$	$Y_{(n_\gamma)}$	$Y_{(n_\gamma+1)}$	$\dots$	$Y_{(N_\gamma-n_\gamma)}$	$Y_{(N_\gamma-n_\gamma+1)}$	$\dots$	$Y_{(N_\gamma}$
$\frac{1}{2}$	1	$\dots$	1	0	$\dots$		$\dots\dots$	$\dots$	0
$\frac{1}{2}$	0	$\dots$	$\dots$	$\dots$	$\dots$	0	1	$\dots$	1

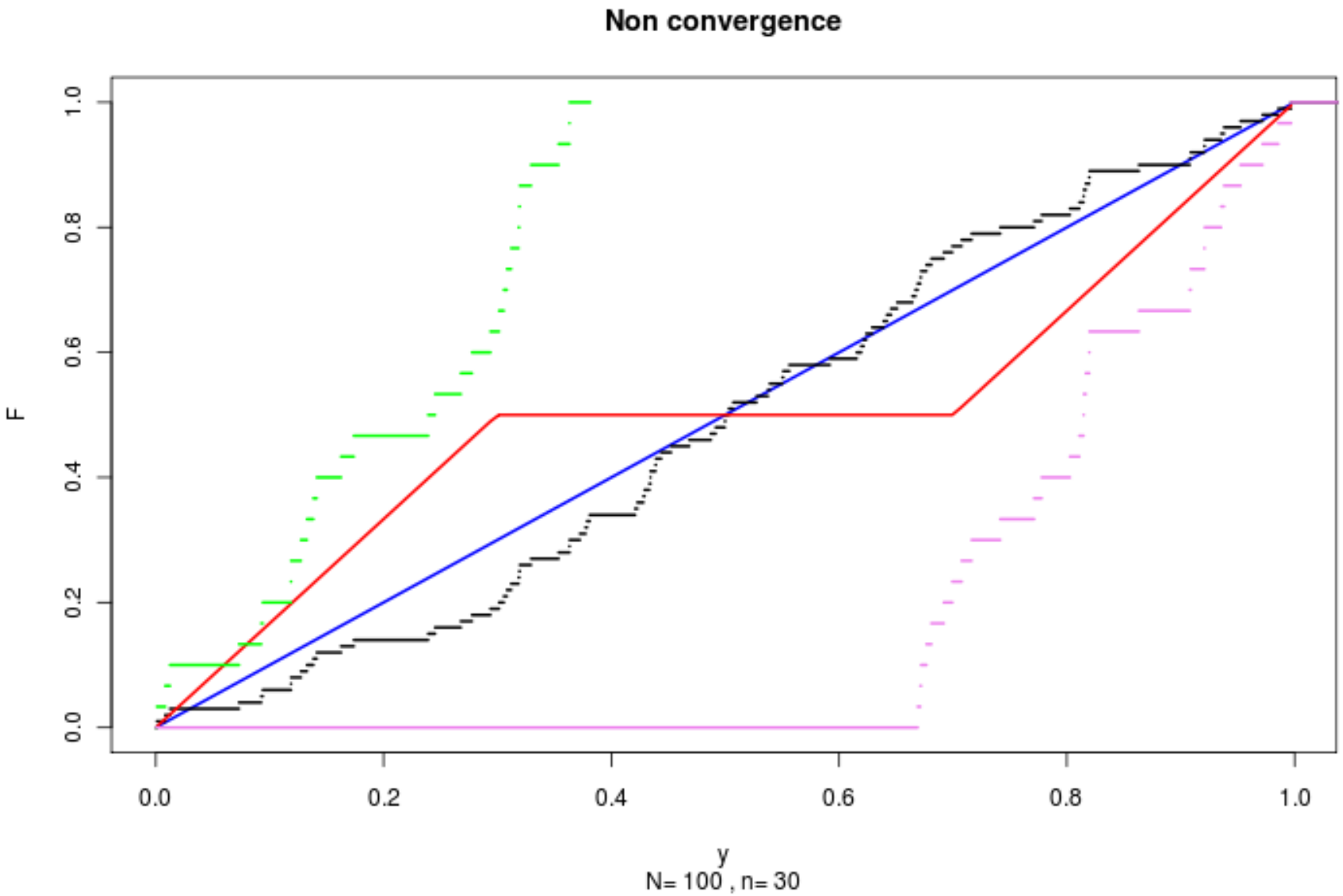
Dans ce cas, A_1, A_2 ne sont pas vérifiées.


Non convergence
F
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
y
N= 100 , n= 30


Non convergence
F
0.0
0.2
0.4
0.6
0.8
1.0
y
N= 100 , n= 30

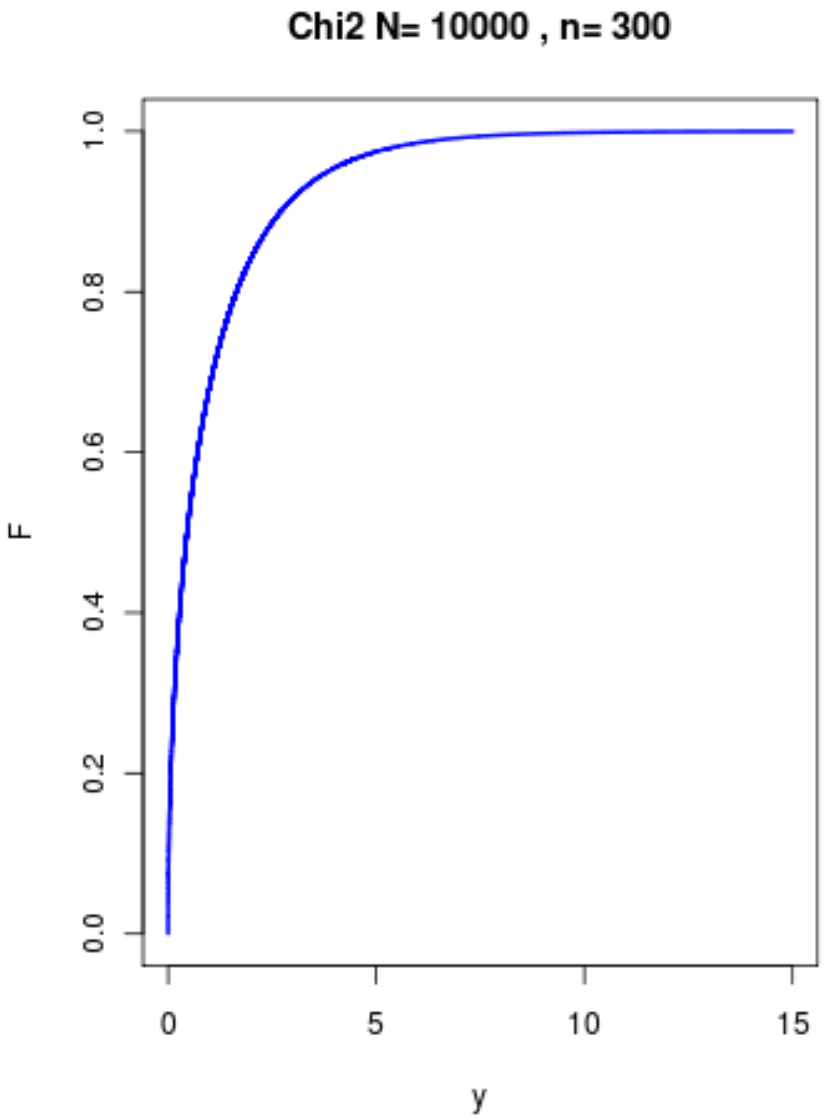

Non convergence
F
0.0
0.2
0.4
0.6
0.8
1.0
y
N= 100 , n= 30

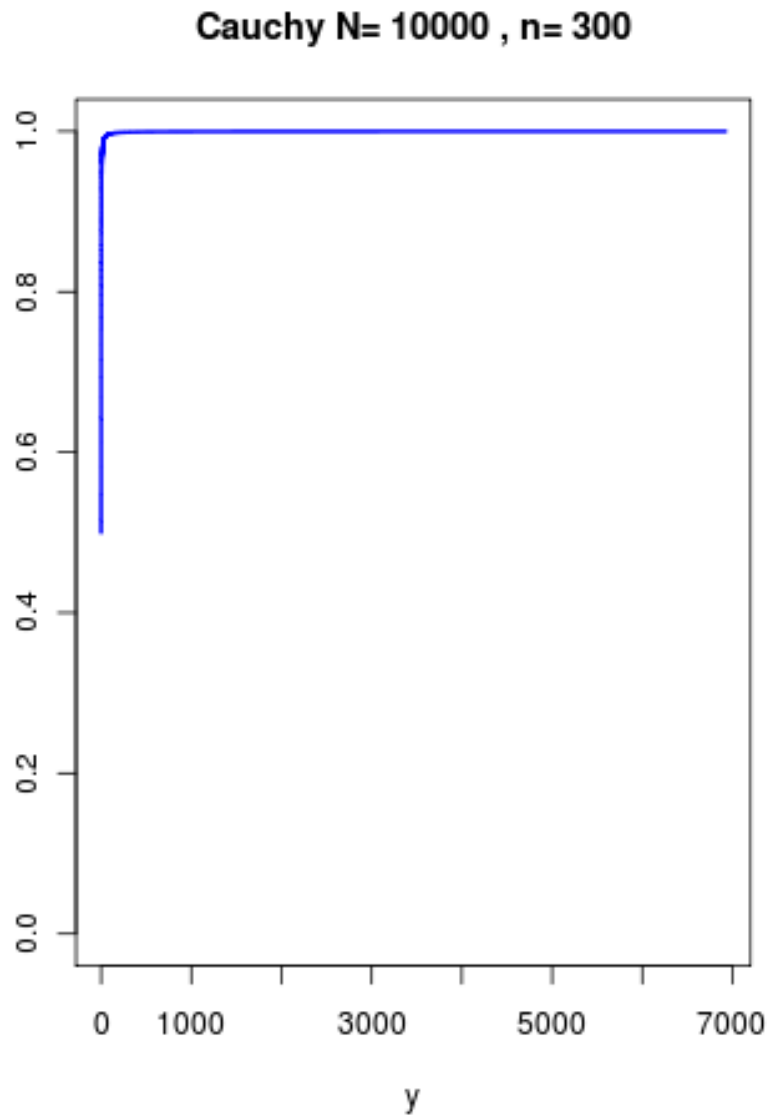

Non convergence
F
0.0
0.2
0.4
0.6
0.8
1.0
y
N= 100 , n= 30

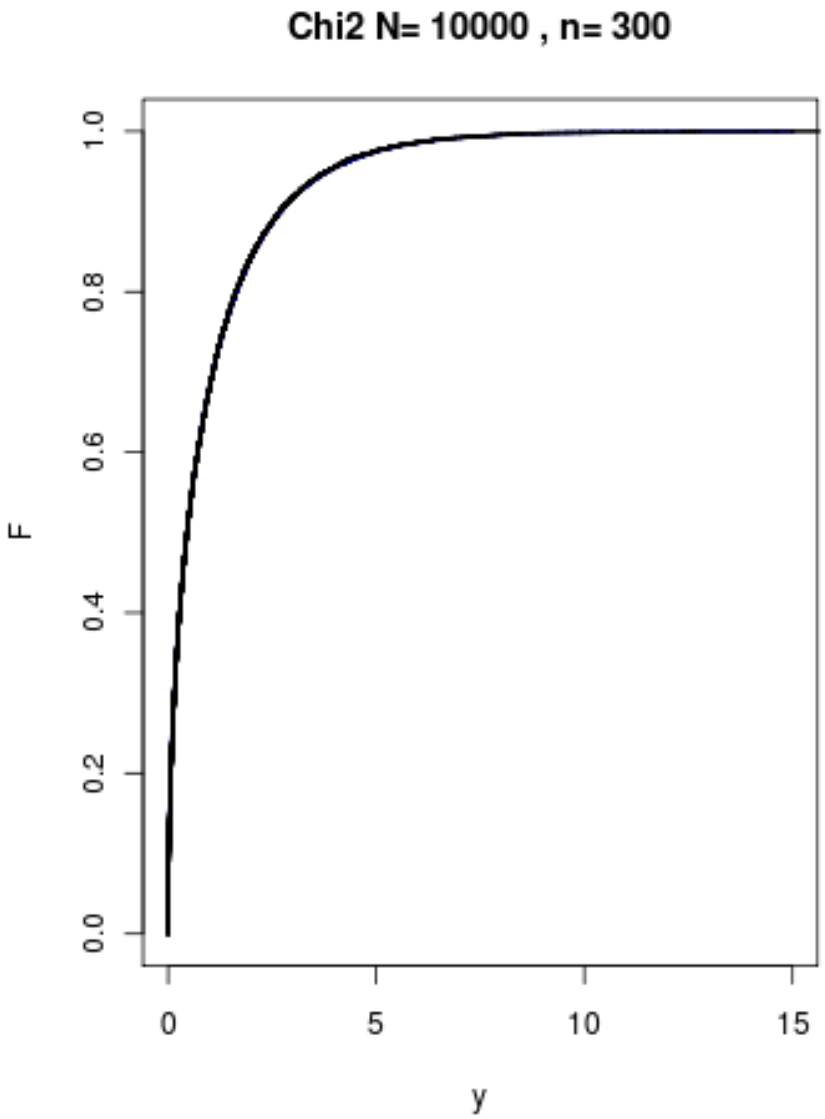

Non convergence
F
0.0
0.2
0.4
0.6
0.8
1.0
y
N= 100 , n= 30

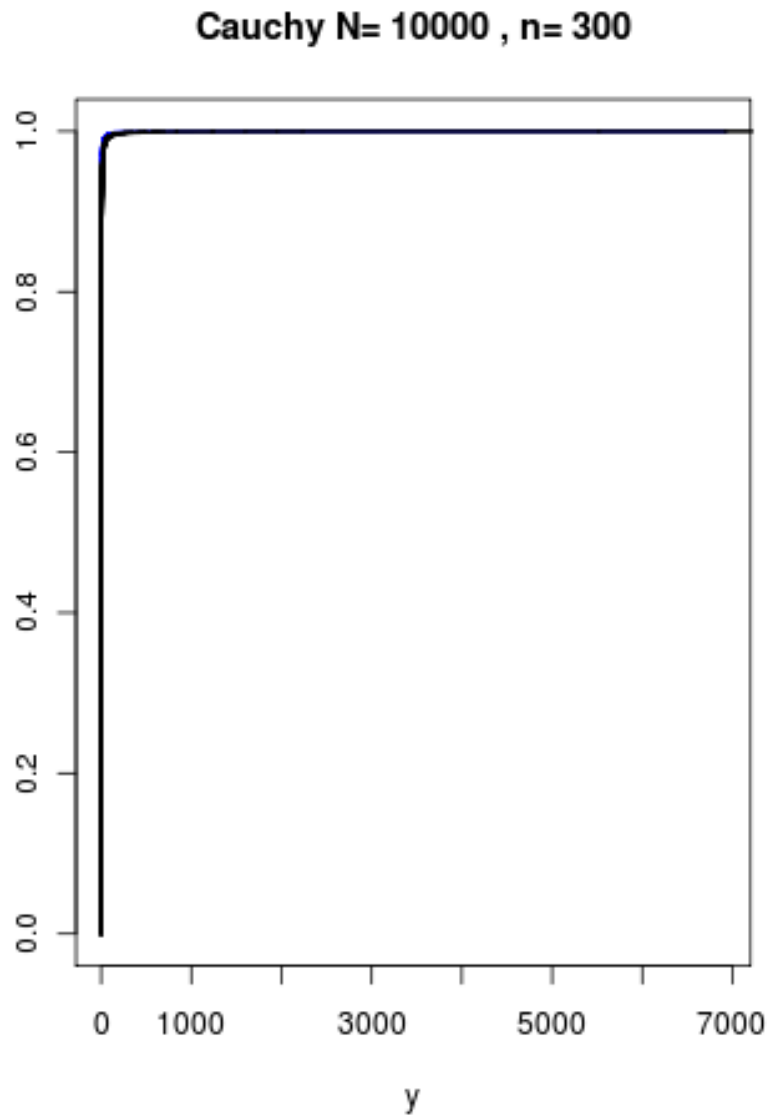
Sondage avec remise et probabilités proportionnelles à la taille

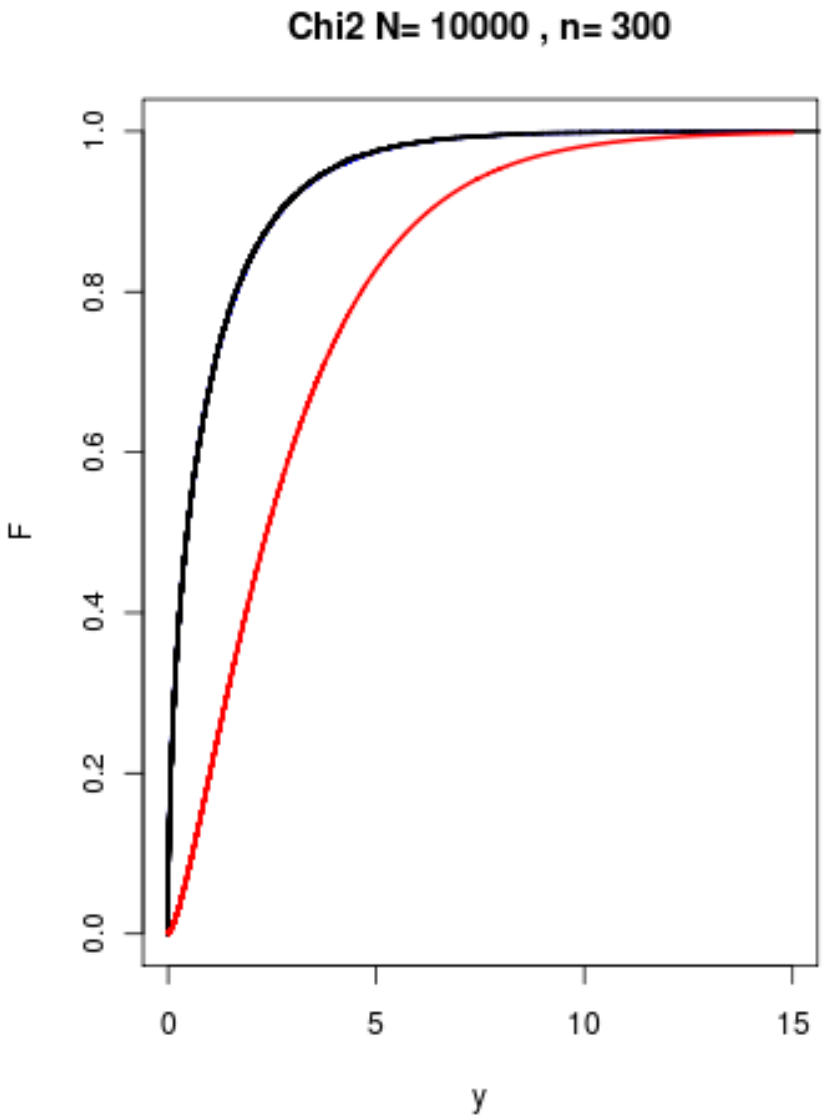
- On considère un tirage avec remise de taille n_γ avec tirages indépendants et probabilités proportionnelles à la taille : $p_\gamma = \frac{Y_\gamma}{\sum_{k=1}^{N_\gamma} Y_{\gamma k}}$, Alors A_0, A_1, A_2 sont vérifiées si $\lim \frac{n_\gamma}{N_\gamma} > 0$ et $Y \in L_6$
- Sans ces conditions sur les moments, A_0, A_1, A_2 peuvent ne pas être vérifiées.

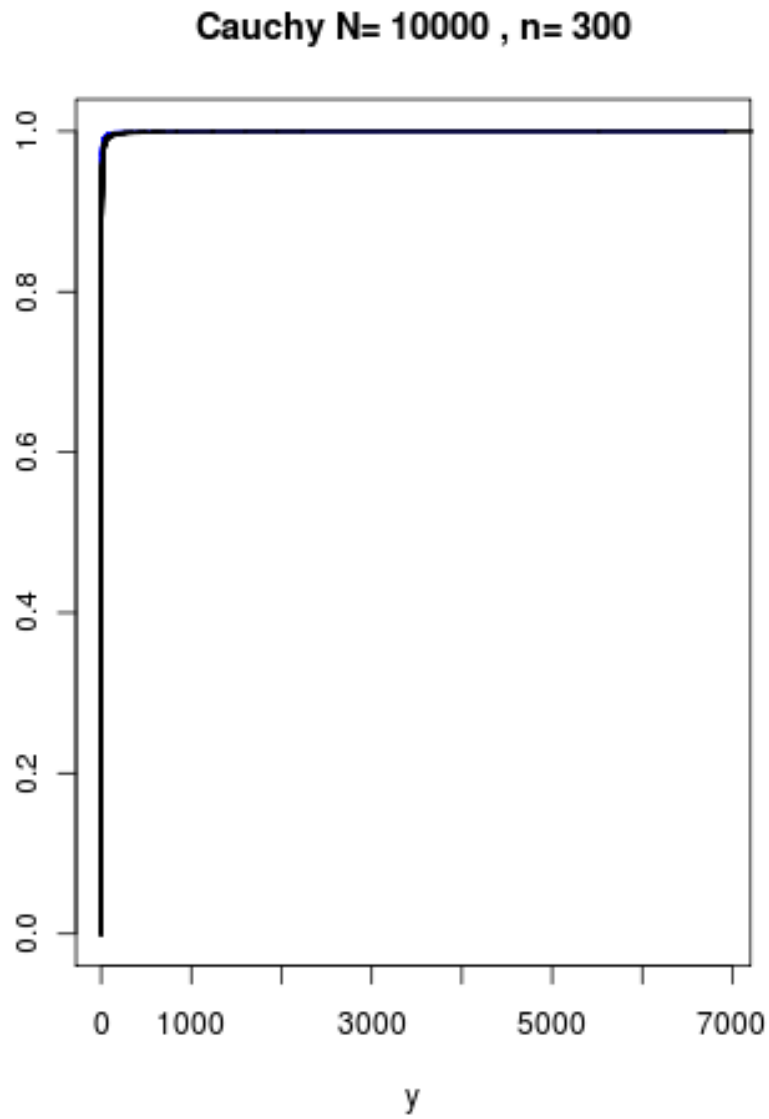

Chi2 N= 10000 , n= 300
F
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
y

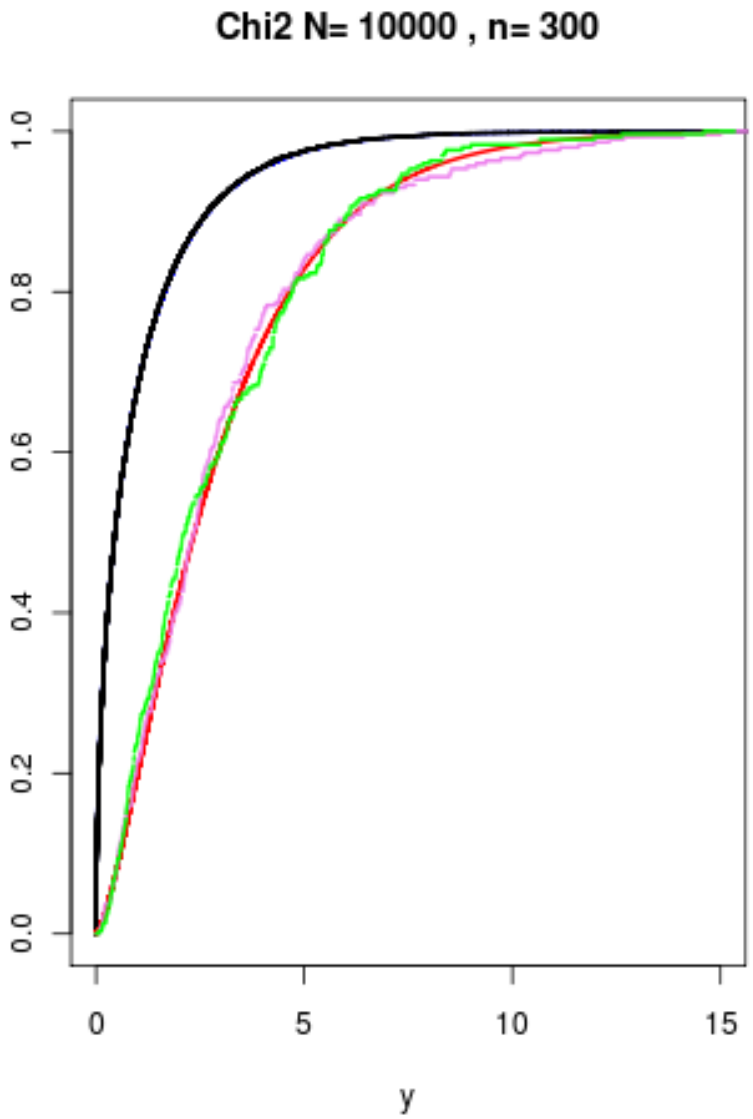

Cauchy N= 10000 , n= 300
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
3000
5000
7000
y

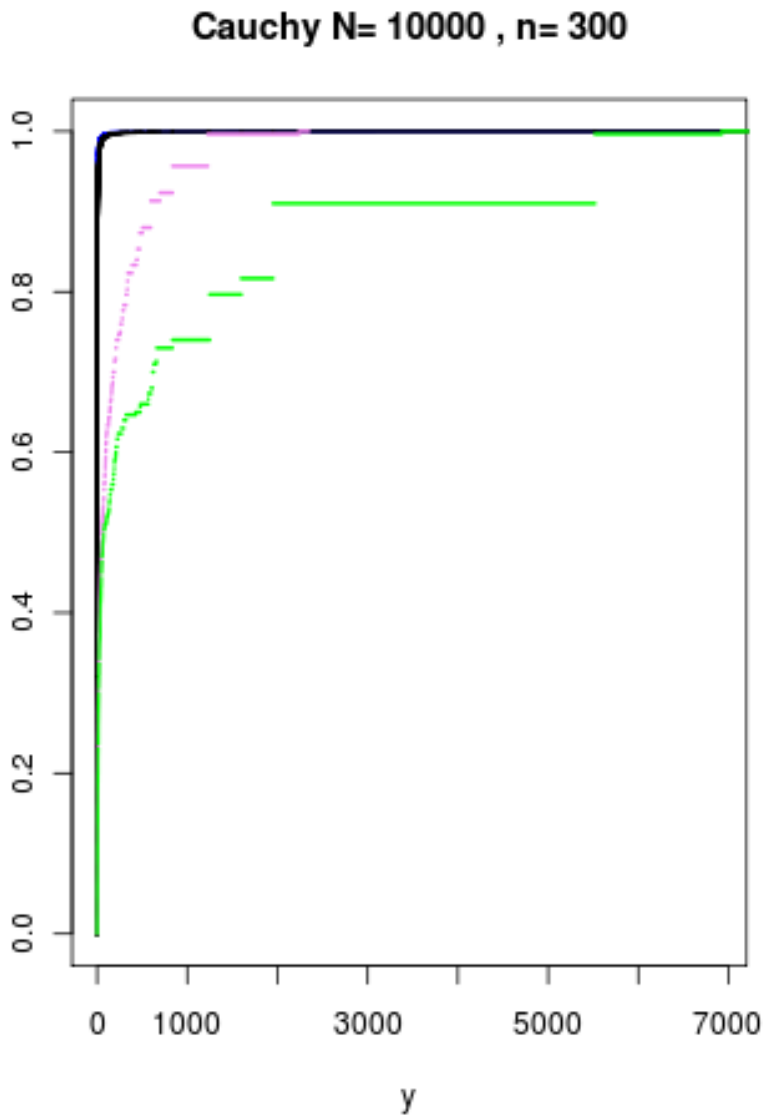

Chi2 N= 10000 , n= 300
F
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
y


Cauchy N= 10000 , n= 300
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
3000
5000
7000
y


Chi2 N= 10000 , n= 300
F
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
y


Cauchy N= 10000 , n= 300
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
3000
5000
7000
y


Chi2 N= 10000 , n= 300
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
y


Cauchy N= 10000 , n= 300
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
3000
5000
7000
y

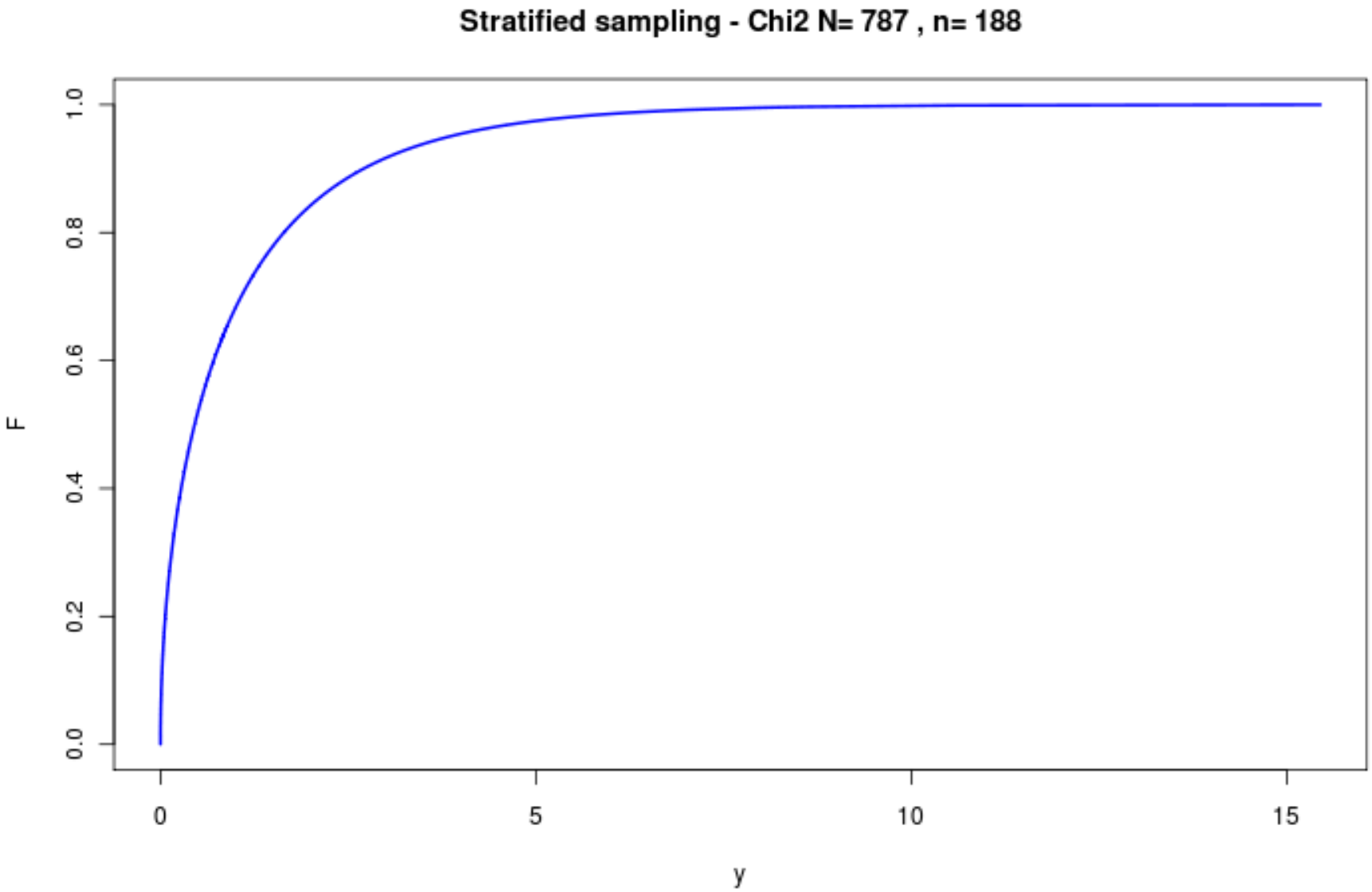
Stratification basée sur Y

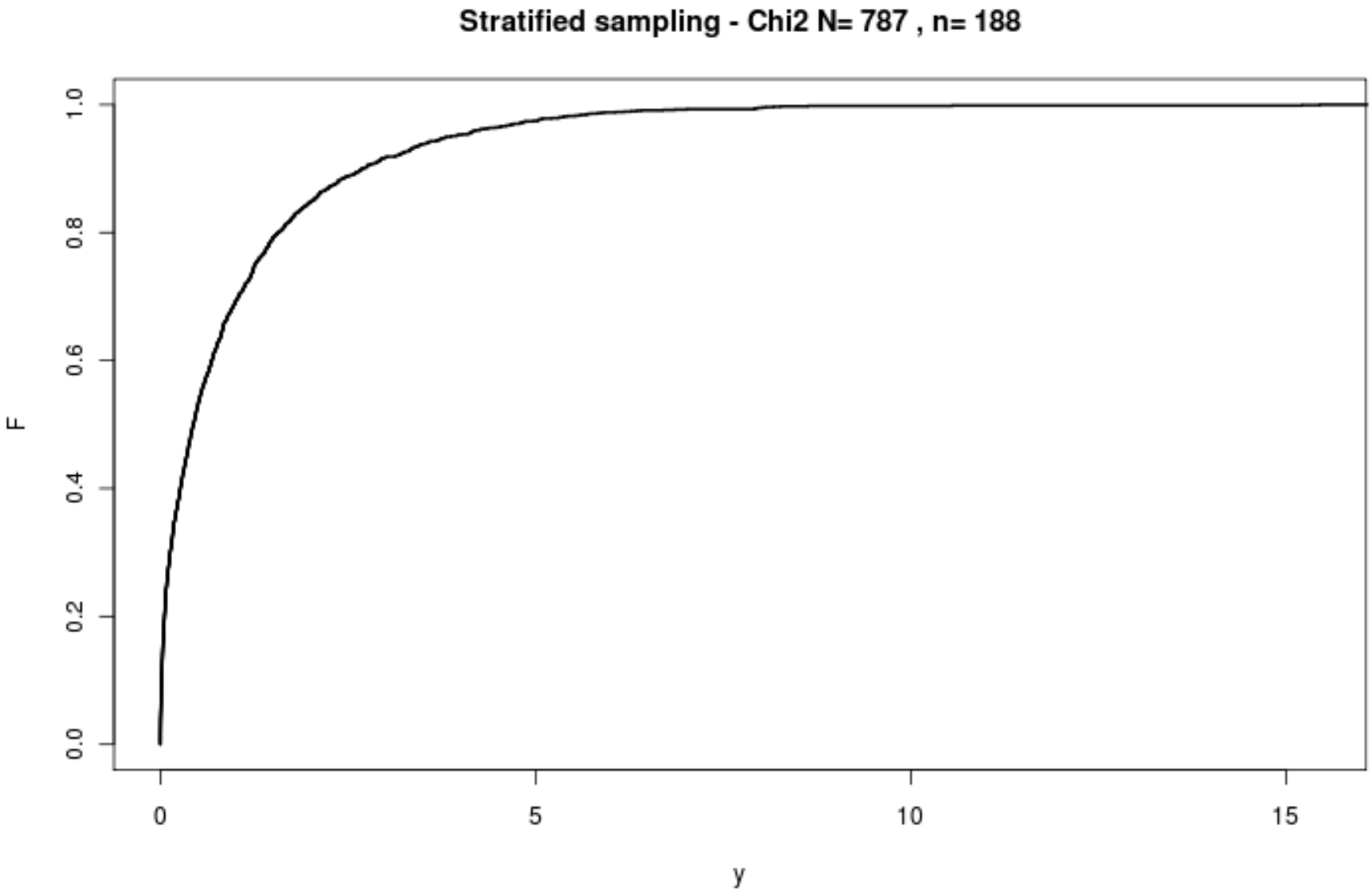
Pour $\gamma \in \mathbb{N}$, la population est stratifiée par tranches de valeurs de Y en H_γ strates. La strate h de taille $N_{\gamma h}$ est

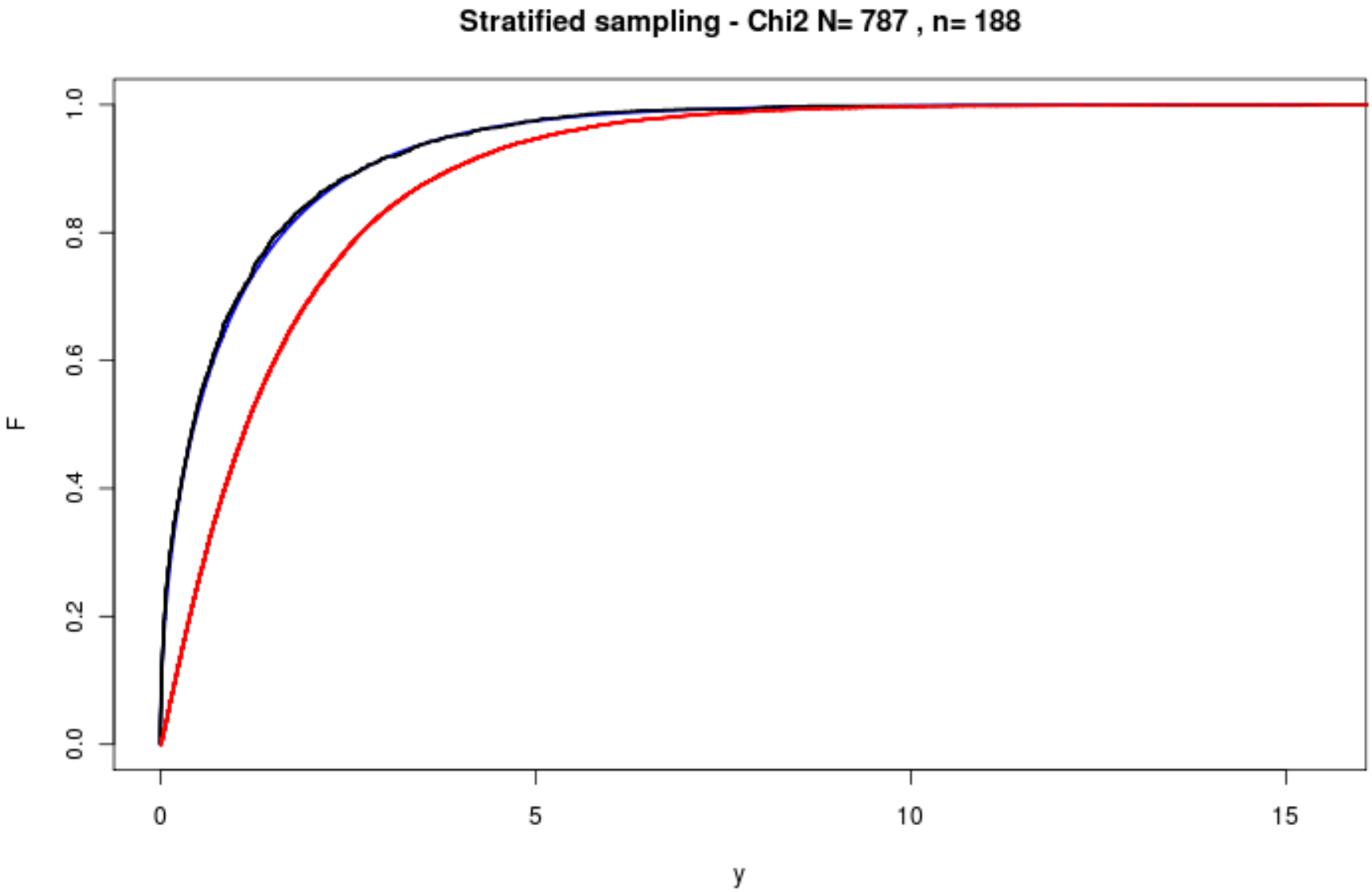
$$\{k \in [\![1, N]\!] | Y_{(\sum_{h'<h} N_{\gamma h'})} < Y_k \leq Y_{(\sum_{h' \leq h} N_{\gamma h'})}\}$$

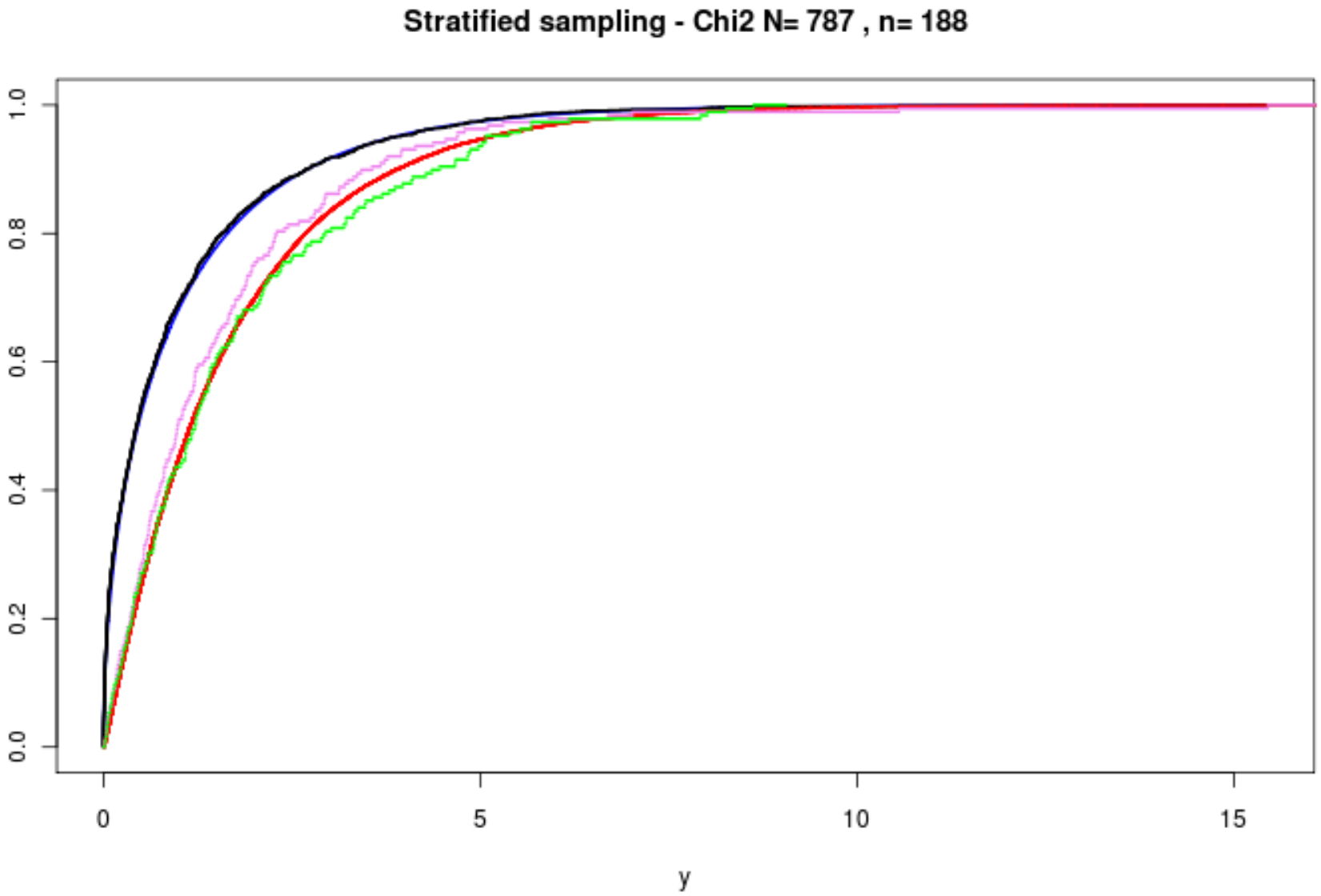
Si $\lim_\gamma \sum_{h=1}^{H_\gamma} \frac{n_{\gamma h}}{N_{\gamma h}} \mathbb{1}_{\left]\frac{\sum_{h'<h} N_{\gamma h'}}{N_\gamma}, \frac{\sum_{h' \leq h} N_{\gamma h'}}{N_\gamma}\right[}$ exists, alors A_0, A_1, A_2 sont vérifiées.

On considère désormais un sondage stratifié avec une allocation $n_{\gamma h}$ croissante avec h.


Stratified sampling - Chi2 N= 787 , n= 188
F
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
y


Stratified sampling - Chi2 N= 787 , n= 188
F
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
y


Stratified sampling - Chi2 N= 787 , n= 188
F
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
y


Stratified sampling - Chi2 N= 787 , n= 188
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
y

Résumé et discussion

Résumé:

- Conditions simples et vérifiables sur la suite de plans de sondage.
- Résultats de convergence uniforme de la fonction de répartition.
- Exemples

Travaux en cours :

- Autres propriétés asymptotiques de statistiques du même type dans le même cadre.
- Normalité asymptotique.


Danny Pfeffermann, Abba M. Krieger, and Yosef Rinott.
Parametric distributions of complex survey data under informative probability sampling.
Statist. Sinica, 8(4):1087–1114, 1998.


Donald B. Rubin.
Inference and missing data.
Biometrika, 63(3):581–592, 1976.


R. A. Sugden and T. M. F. Smith.
Ignorable and informative designs in survey sampling inference.
Biometrika, 71(3):495–506, 1984.