Y-a-t'il un effet de sélection dans la collecte annuelle multimode du recensement ?
Évaluer l'impact de l'endogénéité sans protocole adapté

Loreline Court Simon Quantin (Jeanne Pages)

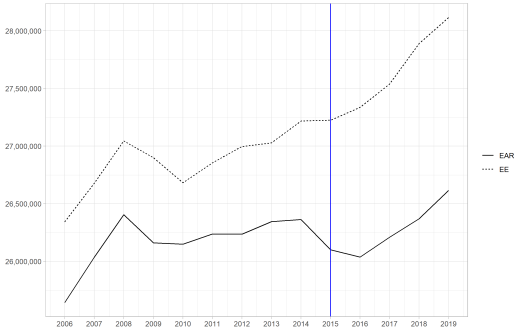
Insee - DMCSI (- DMTR)

30 juin 2023

Introduction

Le nombre de personnes en emploi dans les Enquêtes Annuelles de Recensement (EAR) sont souvent comparées aux estimations annuelles d'emploi (EE) réalisées à partir de plusieurs sources administratives qui sont la référence.

1. Jusqu'en 2014, emploi dans les EAR et les EE présentent une différence constante (environ 700 000 emploi),
2. À partir de 2015, introduction de la collecte multimode dans les EAR : papier et internet,
3. À partir de 2015, l'écart s'est creusé pour atteindre 1 500 000 en 2019.



Peut-on prouver que« l'introduction du multimode » a eu un **effect causal** sur les estimations d'emploi à partir des EAR ?

Introduction

Oui, si l'on considère que l'écart avant 2014 serait resté le même en l'absence de multimode $\longrightarrow$ stratégie de différence de différences

	EAR	EE	Différence
Écart moyen entre 2015 et 2019	26 267 000 (168 000)	27 621 000 (104 000)	1 354 000 (197 000)
Écart moyen entre 2006 et 2014	26 175 000 (77 000)	26 861 000 (87 000)	686 000 (116 000)
Différence	92 000 (130 000)	760 000 (189 000)	668 000 (221 000)

Note : écart-types entre parenthèses, nombres arrondis au millier près.

L'« introduction du multimode » conduit à un accroissement de l'écart de l'ordre de 700 000 emplois.

Il est cependant nécessaire de comprendre les mécanismes qui sous-tendent cet effet. Nous allons d'abord étudier **l'impact du mode de collecte sur la non-réponse, puis questionner l'effet de sélection qui pourrait en résulter**.

Les enquêtes annuelles de recensement

En France, chaque année, le recensement est basé sur une enquête annuelle réalisée auprès d'un **échantillon de logements** et couvrant successivement tous les territoires communaux sur une période de cinq ans.

Pour chaque logement, a minima **deux formulaires sont à remplir**, un pour le logement et un pour chaque habitant.

- questionnaire logement : caractéristiques du ménage (**composition**, **taille**, **liens familiaux**, nombre de voitures, etc.) et du logement lui-même (type, année de construction, taille, type d'occupation, logement social, etc.)

- questionnaire individuel : sexe, âge, statut marital, lieu de naissance, nationalité, lieu d'études, diplôme, **situation principale vis-à-vis de l'emploi**, catégorie socioprofessionnelle, etc.

Différents types de non-réponses

Comme **le logement est l'unité enquêtée (et non ses occupants)**, plusieurs types de non-réponses :

1. **Non-réponse totale** : **pas de questionnaire logement ni de questionnaire individuel**. Se produit lorsque le logement a changé de statut (par exemple, détruit ou vacant) ou qu'il n'y a aucun signe de vie pendant la période de collecte. $\longrightarrow$ considérée **non-réponse aléatoire** (MAR).

2. Non-réponse partielle : **le questionnaire logement est rempli**, mais toujours deux raisons principales pour les non-observations :

   - Un habitant ne remplit *pas du tout le questionnaire individuel* : **non-réponse individuelle**,

   - Un habitant répond au questionnaire individuel, mais celui-ci est *incomplet* : **non-réponse partielle à une question**.

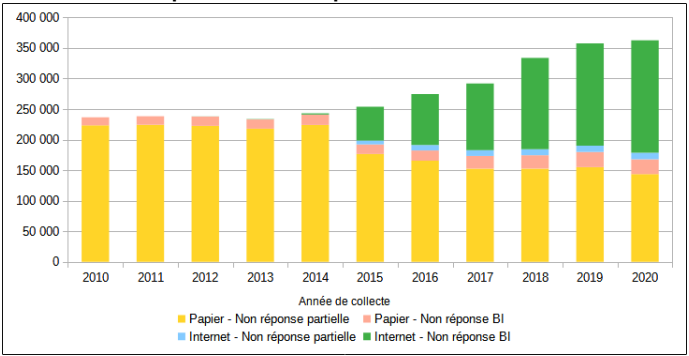
Différents types de non-réponses

Dans chaque logement, **tous les habitants doivent répondre avec le même mode**. (remarque : du point de vue formel, l'indépendance des unités enquêtées n'est donc valide qu'au niveau logement).

Comment internet pourrait impacter la participation ?
Répondre par internet pourrait réduire la non-réponse partielle :
Bien que les questions soient les mêmes quel que soit le mode de collecte, sur internet, un message rouge apparaît en cas de non-réponse (bien que la réponse ne soit pas obligatoire).

Répondre par internet pourrait accroître la non-réponse individuelle :
L'implémentation informatique de l'enquête sur internet permet, à la fin du questionnaire logement, d'envoyer sa réponse à l'enquête sans remplir (tous) les formulaires individuels associés. Cela est pratiquement impossible avec la collecte papier, car les agents recenseurs doivent vérifier que tous les questionnaires individuels ont été (au moins partiellement) remplis avant de les accepter.

Différents types de non-réponse au questionnaire individuel



Depuis l'introduction du multimode, plus il y a de répondants par internet, plus il y a de non-réponses.

La non-réponse individuelle (tout le questionnaire individuel non rempli) est presque le seul type de non-réponse des répondants par internet.

La non-réponse partielle (ici à la question sur l'emploi) est presque le seul type de non-réponse des répondants papier.

Méthodologie - Principes

Ces chiffres reflètent-ils :

- un effet de la conception du questionnaire internet sur la non-réponse ? **et/ou**

- un biais de sélection (les enquêtés sur internet diffèrent des enquêtés sur papier, ce qui entraîne différents types de non-réponse),

Y répondre revient à tester l'hypothèse (nulle) de l'absence d'effet de la conception du questionnaire internet sur la non-réponse, en comparant les taux de participation entre les logements répondant sur internet et sur papier.

En l'absence d'affectation aléatoire à un mode, les logements peuvent ne pas être comparables au moment de répondre à l'enquête
$\longrightarrow$ appariement sur les caractéristiques observées pour éliminer le biais de sélection
(avec les réponses au questionnaire logement, notamment la composition du ménage)

Cependant, les logements appariés pourraient ne pas être comparables en présence d'hétérogénéité inobservée.

Méthodologie - Principe

Si après appariement, les logements ne sont pas comparables - précisément parce qu'il existe une caractéristique inobservée qui les distingue - alors la corrélation éventuellement mise en évidence, ici entre mode de collecte et participation, ne vaut pas causalité.

Cependant, « **de quelle ampleur doit être le biais de sélection non observé résiduel, pour disqualifier toute causalité dans la corrélation mise en évidence** » ? C'est une réponse à cette question que propose le modèle d'analyse de sensibilité de Rosenbaum (2002).

Formellement, le modèle s'appuie sur les données pour **quantifier l'incertitude**, liée à une caractéristique inobservée, **sur nos conclusions**, tout comme un intervalle de confiance témoigne de l'incertitude liée à l'échantillonnage.

Comment ? On va considérer que malgré l'appariement, l'un des logements a, par exemple, *toujours* trois fois plus de chances de répondre par internet que l'autre. Si nos conclusions ne changent pas malgré un tel écart par rapport à l'affectation aléatoire [1], alors ces conclusions seront dites « insensibles » à l'existence d'un biais de sélection d'une telle ampleur.

1. conditionnellement aux observables retenues

Méthodologie - Le modèle d'analyse de sensibilité

Formellement, comme il existe de l'hétérogénéité inobservée, les probabilités de répondre par internet, π_{si}, de chaque logement i dans une paire s diffèrent et sont inconnues.

Le modèle de sensibilité pose que cette différence (d'odds-ratio) est encadré par un facteur $\Gamma \geq 1$.

$$\frac{1}{\Gamma} \leq \frac{\pi_{si}/(1-\pi_{si})}{\pi_{sj}/(1-\pi_{sj})} \leq \Gamma \tag{1}$$

Ou, pour le dire autrement, il encadre la probabilité, θ_s, d'observer au sein d'une paire, tel logement, répondre par internet :

$$\frac{1}{1+\Gamma} \leq \theta_s \leq \frac{\Gamma}{1+\Gamma} \tag{2}$$

Exemple : Quand $\Gamma = 1$, les π_{si} sont identiques, il y a une chance sur deux d'observer chaque logement répondre par internet au sein d'une paire ($\theta_s = \frac{1}{2}$) : cas de l'affectation aléatoire.

Quand $\Gamma = 9$, les π_{si} diffèrent, il peut y avoir jusqu'à $90\ \%$ de chances que nous observions le logement 1 répondre sur internet, malgré notre appariement exact ($\theta_s \in [0.10; 0.90]$) .

Méthodologie - Test de l'hypothèse nulle d'absence d'effet causal

Ceci posé, comment tester l'hypothèse nulle, d'absence d'effet du mode de collecte sur la participation, dans le cadre du modèle d'analyse de sensibilité ?

Un test repose sur la distribution de la statistique considérée (cf. Annexe).

Cette distribution dépend de la probabilité (π_{si}) de chaque logement de répondre sur internet, probabilité qui nous est inconnue, en raison de l'existence d'une caractéristique non observée : donc, la distribution de la statistique de test est, elle aussi, inconnue.

Comment faire ? Cette distribution peut être encadrée par celles de deux autres statistiques du même test mais dont les paramètres ne dépendent que de Γ (et $\frac{1}{\Gamma}$) , ce qui nous permet d'en tirer des inférences (Rosenbaum, 2002), en encadrant la P-value du test considéré. (cf. Annexe)

Méthodologie - Démarche

Dans chaque paire, on considère que l'un des deux logements a, au maximum, Γ plus de chances de répondre par internet.

En faisant varier Γ, nous pouvons discuter de la sensibilité des résultats à la présence d'une caractéristique non observée.

Par exemple, nous pouvons trouver la valeur maximale de Γ, c'est-à-dire la taille du biais de sélection qui devrait subsister après l'appariement, pour accepter l'hypothèse nulle d'absence d'effet du mode de collecte par internet sur la non-réponse.

Comment lire nos résultats ?

- Si l'on considère que $\Gamma = 1$, il y a une chance sur deux que, après l'appariement exact, le logement 1 ait répondu sur internet et non le logement 2.
- En d'autres termes, les logements qui semblent comparables sont effectivement comparables (pas d'hétérogénéité inobservée).

Γ	Intervalle des valeurs possible de θ_s		Non-réponse Individuelle	Partielle
1	0.50	0.50		

Comment lire nos résultats ?

- Si l'on considère que $\Gamma = 1$, il y a une chance sur deux que, après l'appariement exact, le logement 1 ait répondu sur internet et non le logement 2.
- En d'autres termes, les logements qui semblent comparables sont effectivement comparables (pas d'hétérogénéité inobservée).

Γ	Intervalle des valeurs possible de θ_s		Non-réponse Individuelle	Partielle
1	0.50	0.50	< 0.001	

1. Dans ce cas, on rejette l'hypothèse d'absence d'effet du mode de collecte sur la non-réponse individuelle : répondre par internet engendre de la non-réponse individuelle pour certains ménages,

Comment lire nos résultats ?

- Si l'on considère que $\Gamma = 1$, il y a une chance sur deux que, après l'appariement exact, le logement 1 ait répondu sur internet et non le logement 2.
- En d'autres termes, les logements qui semblent comparables sont effectivement comparables (pas d'hétérogénéité inobservée).

Γ	Intervalle des valeurs possible de θ_s		Non-réponse Individuelle	Non-réponse Partielle
1	0.50	0.50	< 0.001	< 0.001

1. Dans ce cas, on rejette l'hypothèse d'absence d'effet du mode de collecte sur la non-réponse individuelle : répondre par internet engendre de la non-réponse individuelle pour certains ménages,
2. De même, répondre par internet réduit la non-réponse partielle pour certains ménages.

Comment lire nos résultats ?

- Quand $\Gamma = 3$, on autorise que malgré l'appariement, il y ait 75 % de chances que le logement 1 ait répondu sur internet et non le logement 2.
- En d'autres termes, des logements qui semblent comparables ne le sont pas, en raison de l'existence d'une covariable non observée.

Γ	Intervalle des valeurs possible de θ_s		Non-réponse Individuelle	Non-réponse Partielle
1	0.50	0.50	< 0.001	< 0.001
3	0.25	0.75	< 0.001	< 0.001

1. Dans ce cas, on rejette **toujours** l'hypothèse d'absence d'effet du mode de collecte sur la non-réponse individuelle : internet engendre **toujours** de la non-réponse individuelle pour certains ménages,
2. De même, internet réduit **toujours** la non-réponse partielle pour certains ménages.

P-value maximale associée au test $H_0 : \delta = 0$ en fonction de Γ

Γ	Intervalle des valeurs possibles de θ_s		Non-réponse (i)		(ii)	
			Individuelle totale	Individuelle	Partielle totale	Partielle
1	0.50	0.50	< 0.001	< 0.001	< 0.001	< 0.001
3	0.25	0.75	< 0.001	< 0.001	< 0.001	< 0.001
5	0.17	0.83	< 0.001	< 0.001	< 0.001	< 0.001
7	0.13	0.88	< 0.001	< 0.001	< 0.001	< 0.001
9	0.10	0.90	< 0.001	< 0.001	0.981	< 0.001
11	0.08	0.92	< 0.001	0.301		< 0.001
13	0.07	0.93	< 0.001			< 0.001
15	0.06	0.94	< 0.001			
17	0.06	0.94	0.013			
19	0.05	0.95	0.968			
21	0.05	0.95				
23	0.04	0.96				

Note : P-value associée à la statistique de test de McNemar. Une valeur manquante indique une P-value > 0.999.

Les valeur élevées de Γ suggèrent que la conclusion d'un effet causal du mode de collecte sur la non-réponse individuelle et partielle est insensible à la présence d'hétérogénéité inobservé.

Retour à la question initiale

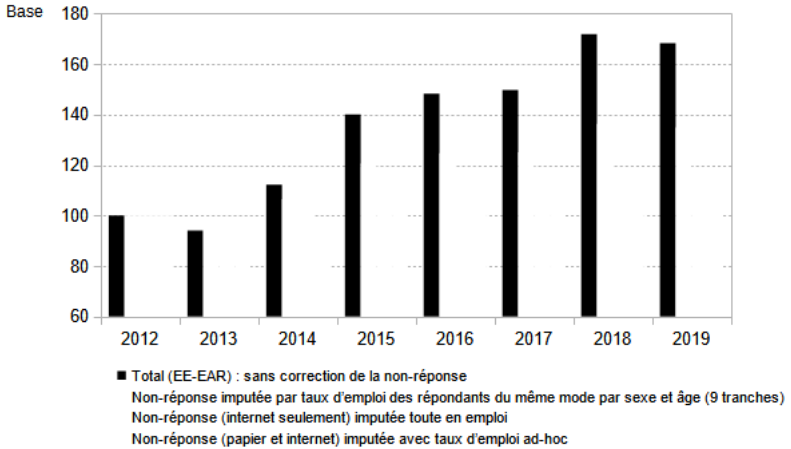
Répondre par internet engendre de la non-réponse individuelle *pour certains ménages* et réduit la non-réponse partielle *pour certains ménages* $\longrightarrow$ le mode de collecte impacte le **type** de non-réponse.

L'existence d'un tel effet questionne tout d'abord le mécanisme actuel de correction de la non-réponse par imputation. En effet, actuellement :

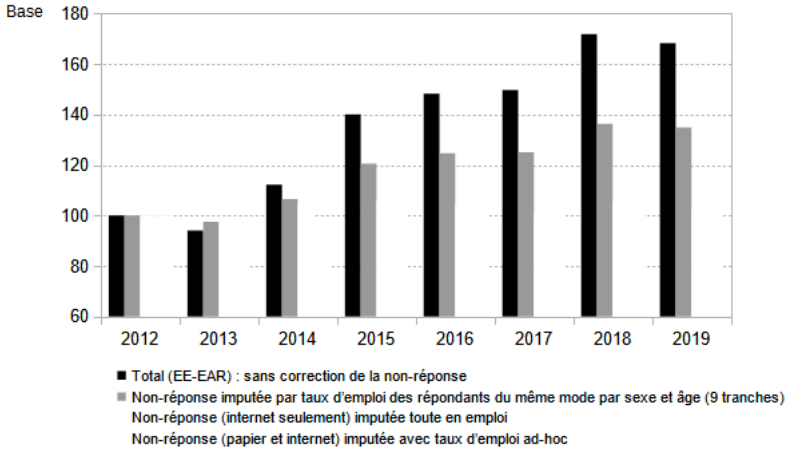
- la non-réponse partielle à la question sur l'emploi est imputée en utilisant d'autres questions,
- la non-réponse individuelle est imputée uniquement à partir des répondants (imputation *hotdeck* par sexe et âge), et sans tenir compte du mode de collecte.
- Comme les répondants papier se déclarent moins souvent en emploi et sont de moins en moins nombreux, le système d'imputation des données manquantes pourrait ne pas imputer assez de non-répondants internet en emploi.

Cependant, cet effet causal **ne dit rien sur qui sont ces ménages**, i.e. sur l'effet de sélection lié au mode de collecte.

Évolution de l'écart entre emploi déclaré et EE, avant imputation

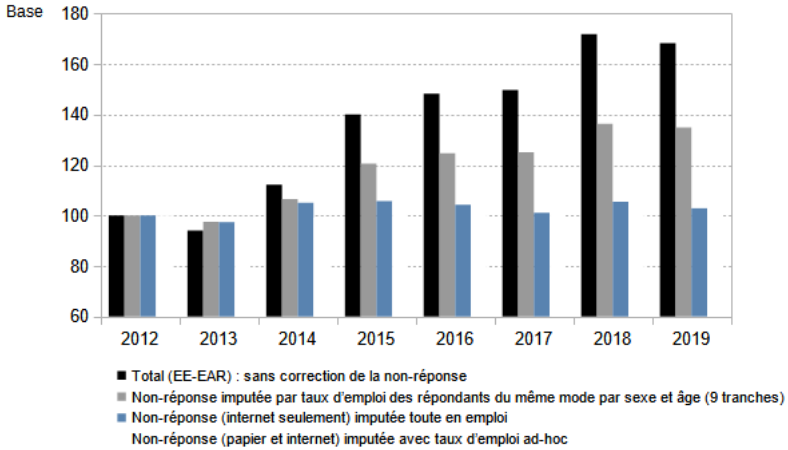


Un effet de sélection ?

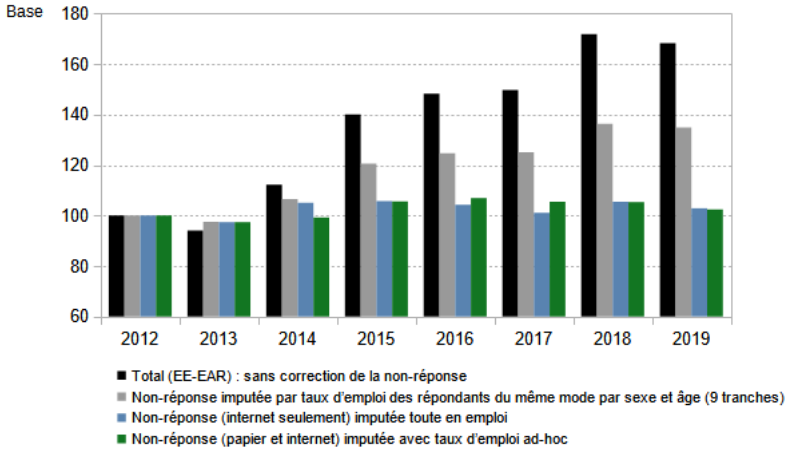


Un écart subsiste en imputant les taux d'emploi des répondants, même en distinguant par mode de collecte.

Des non-répondants par internet en emploi ?



Des non-répondants en emploi ?



Conclusion

On a mis en évidence l'impact du mode de collecte sur les types de non-réponse, un effet qui fragilise la démarche actuelle de correction de la non-réponse.

Cet effet causal sur la participation se double d'un effet de sélection : les non-répondants seraient plus souvent en emploi que les répondants.

Du point de vue méthodologique, la démarche présentée ici est utilisable pour mettre en évidence un effet de mesure - en l'absence de protocole dédié -

- en prenant comme variable expliquée une indicatrice (oui/non) à la question posée,
- et en comparant les répondants avec un modèle d'analyse de sensibilité, pour tenir compte de l'existence d'un effet de sélection endogène/inobservé.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer Series in Statistics.

Annexe A : Expérience aléatoire et test de l'absence d'effet causal

On veut tester l'existence d'un effet du mode de collecte sur la probabilité de non-réponse $H_0 : \delta_i = 0$.

Toute statistique de test est une fonction de mode de collecte $\mathbf{Z}$ et des réponses observées $\mathbf{R}$: $t(\mathbf{Z}, \mathbf{R})$.

Dans le cas d'une expérience aléatoire, $\mathbf{Z}$ a une distribution connue. Par exemple, si on sélectionne au hasard m logements, parmi n, qui doivent répondre par internet et $n-m$ qui doivent répondre par papier, $\mathbb{P}(\mathbf{Z} = \mathbf{z}) = \frac{1}{|\mathcal{Z}|}$ avec $|\mathcal{Z}| = \binom{n}{m}$.

En l'absence d'effet de mode (sous H_0), $\mathbf{R} = \mathbf{r_C} = \mathbf{r_T}$ est fixe : la participation ou non ne dépend pas du mode de collecte.

Donc $t(\mathbf{Z}, \mathbf{r}_C)$ est une fonction de grandeurs fixes et d'une variable aléatoire $\mathbf{Z}$, **seule source d'aléa et de distribution connue**. La P-value associée au test est donnée par :

$$\mathbb{P}\big\{t(\mathbf{Z},\mathbf{r}) \geq T\big\} = \sum_{\mathbf{z}\in\mathcal{Z}} \mathbb{1}_{\{t(\mathbf{z},\mathbf{r})\geq T\}} \mathbb{P}(\mathbf{Z} = \mathbf{z})$$

Calcul direct puisque $\mathbb{P}(\mathbf{Z} = \mathbf{z})$ est connue (mais fastidieux si $|\mathcal{Z}|$ est grand).

Annexe A : Expérience aléatoire et test de l'absence d'effet causal

Considérons $t(\mathbf{Z},\mathbf{r})$, nombre de non-répondants par internet et notre exemple d'affectation aléatoire uniforme.

Fisher (1935) a montré que la P-value précédente peut être déterminée en considérant le tableau de contingence suivant conditionnellement aux totaux marginaux [2] :

- R_+ est fixe sous H_0.
- m et n le sont par le protocole aléatoire d'assignation

	Non-réponse $R_i = 1$	Réponse $R_i = 0$	Total
Internet $Z_i = 1$	$T = \sum Z_i r_{Ci}$	$-$	m
Papier $Z_i = 0$	$-$	$-$	$n - m$
Total	$R_+ = \sum r_{Ci}$	$n - R_+$	n

Formellement, il s'agit du test exact de Fisher. Il est « exact » car $\mathbb{P}\{t(\mathbf{Z},\mathbf{r}) = T\}$ peut être calculé pour tout T, à partir d'une loi hypergéométrique de paramètres (R_+, m, n) et à partir de là, la P-value associée. retour

2. équivalence avec un test de permutation d'où provient la formule précédente

Annexe B : Analyse de sensibilité

Rosenbaum (2002) montre alors que dans ce cas, la statistique de test $T = t(\mathbf{Z},\mathbf{R})$ vérifie pour Γ fixé et pour tout a :

$$\mathbb{P}(T^+ \geq a) \geq \mathbb{P}(T \geq a \mid \mathbf{m}) \geq \mathbb{P}(T^- \geq a)$$

où les distributions de T^+ et T^- sont connues et ne dépendent que de Γ.
Lorsque $\Gamma = 1$, $\mathbb{P}(T^+ \geq a) = \mathbb{P}(T^- \geq a)$.

Par exemple, sur données appariées par paires, dans le cas d'une variable d'intérêt R binaire, les distributions de T^+ et T^- suivent la distribution de la somme de lois binomiales de probabilité $p^+ = \frac{\Gamma}{(1+\Gamma)}$ et $p^- = \frac{1}{(1+\Gamma)}$ respectivement. retour