SIMULATION D'UN SONDAGE D'OPINION

par Vincent Papillon

Introduction

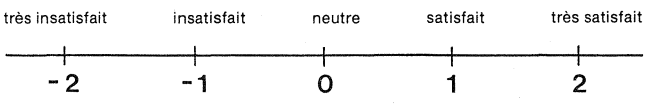
Les mathématiques, en faisant abstraction de la complexité du monde réel et en ne retenant que certaines variables, permettent de construire des modèles qui imitent une partie de la réalité. Le présent article a précisément pour but de montrer comment on peut simuler un mini-sondage d'opinion; cette simulation permet d'engendrer des données fictives conformes à des caractéristiques globales préétablies.[1] Le modèle de simulation proposé ici n'a aucune prétention de réalisme sociologique ou politique; il s'agit d'une simple fiction mathématique dont l'intérêt est purement pédagogique.

Caractéristiques du modèle de simulation

Le modèle qu'on se propose de construire doit imiter des répondants choisis au hasard dans une population imaginaire lors d'un sondage d'opinion. Plus précisément, il faut imiter des personnes qui répondent à un même questionnaire, indépendamment les unes des autres, mais conformément aux caractéristiques globales d'une population donnée. On peut donc commencer la construction du modèle en imaginant un questionnaire. Mathématiquement, cela signifie qu'on doit choisir les **variables** sous lesquelles on veut considérer les répondants. Dans ce sondage fictif, on suppose qu'on veut classifier les répondants selon les caractères suivants:

le sexe **Sx**, l'âge **Ag**,
le niveau de scolarité **Scol**,
le revenu annuel net **Rev**,
le degré de satisfaction à l'endroit du gouvernement **Sat**,
l'opinion sur un projet gouvernemental donné **Opin**.

Cela fait déjà six variables. Lorsque ce modèle *fonctionnera* sur ordinateur (ou autrement), il devra imprimer une ligne d'information pour chaque répondant simulé; sur cette ligne devra apparaître le sexe du répondant (M ou F), son âge, son niveau de scolarité (nombre d'années d'étude à temps complet), son revenu annuel net en milliers de $, son degré de satisfaction sur l'échelle



et aussi son opinion (A: pour B: contre C: indécis). Lorsque le modèle sera en opération, il imprimera selon le format ci-dessous, en précisant ligne par ligne.

répondant no	Sx	Ag	Scol	Rev	Sat	Opin
1						
2						
3						
4						

1037	F	43	17	29	1	A

Dans l'exemple ci-dessus, le répondant no 1037 est de sexe féminin, il est âgé de 43 ans, il a l'équivalent de 17 années d'études à temps complet, il a actuellement un revenu annuel net de 29 000$, il est satisfait du gouvernement et il est en faveur du projet proposé.

Construction du modèle

La construction du modèle consiste à déterminer les règles suivant lesquelles on choisit les valeurs (ou modalités) des six variables pour chaque répondant. Par exemple, pour la variable **Sx**, la règle pourrait consister à lancer une pièce de monnaie équilibrée et à inscrire ***M*** (masculin) si on obtient Pile et ***F*** (féminin) si on obtient Face; dans ce cas, la population visée par le sondage serait composée d'autant d'hommes que de femmes. On peut cependant modifier cette situation selon sa propre fantaisie: pourquoi la proportion des hommes dans cette population imaginaire ne serait-elle pas 0,43 et celle des femmes 0,57? La règle de simulation pour la variable **Sx** pourrait alors devenir:

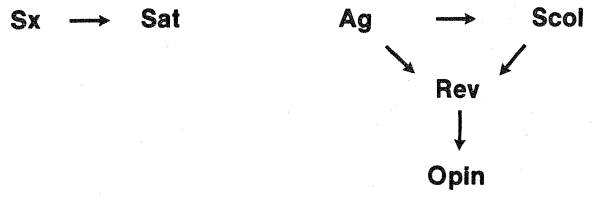
*on choisit un nombre entier au hasard uniformément entre 1 et 100; si le nombre choisi est inférieur ou égal à 43, le sexe du répondant est **M**, autrement c'est **F**.*

Sur un grand nombre de répondants, on peut espérer que cette règle fournirait une proportion de ***M*** voisine de 0,43. On voit que les règles pour le calcul des valeurs des variables correspondent à une image préétablie de la population visée. Il faut donc établir à l'avance les caractéristiques de cette population face aux six variables du modèle: liens de dépendance entre ces variables, distributions statistiques des variables.

Liens de dépendance entre les variables

On pourrait imaginer une population tout-à-fait artificielle dans laquelle, par exemple, la distribution des âges serait très différente chez les hommes et chez les femmes. Dans un tel cas, les variables ***Ag*** et ***Sx*** seraient dépendantes. Il est plus vraisemblable de considérer les variables ***Sx*** et ***Ag*** comme indépendantes. Le

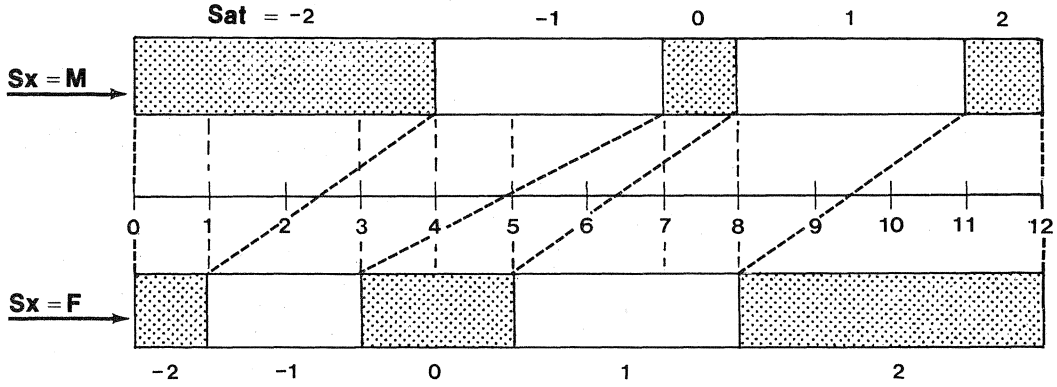
schéma suivant illustre comment les variables sont reliées dans le modèle proposé.



Dans ce schéma, une flèche ***A*** → B indique que la variable ***B*** dépend de la variable ***A***. Par exemple, ***Rev*** dépendra de ***Ag*** et de ***Scol*** mais sera indépendante de ***Sx*** (le choix et la nature des liens de dépendance est arbitraire et guidé seulement par l'usage éventuel des données). Voici, par exemple, comment on peut déterminer graphiquement (et arbitrairement...) la relation de dépendance statistique entre les variables ***Sx*** et ***Sat*** dans l'ensemble de la population visée par le sondage. Selon le diagramme ci-dessous, les hommes sont très satisfaits (***Sat*** = −2) dans une proportion de 4/12, tandis que les femmes sont très insatisfaites dans une proportion de 1/12. La proportion des personnes très insatisfaites dans l'ensemble de la population visée serait alors:

$$\frac{4}{12}(0{,}43) + \frac{1}{12}(0{,}57) \simeq 0{,}19$$

proportion des hommes — proportion des femmes



(échelle)

En pratique, lorsque le sexe d'un répondant a été déterminé (par la règle correspondant à **Sx**), on peut déterminer son degré de satisfaction **Sat** en traduisant le diagramme de dépendance par la règle suivante:

On choisit un nombre entier de 1 à 12 au hasard, uniformément; soit u ce nombre;
si **Sx** = *M et*

si $u \leq 4$, *alors* **Sat** = -2
si $5 \leq u \leq 7$, *alors* **Sat** = -1
si $u = 8$, *alors* **Sat** = 0
si $9 \leq u \leq 11$, *alors* **Sat** = 1
si $u = 12$, *alors* **Sat** = 2

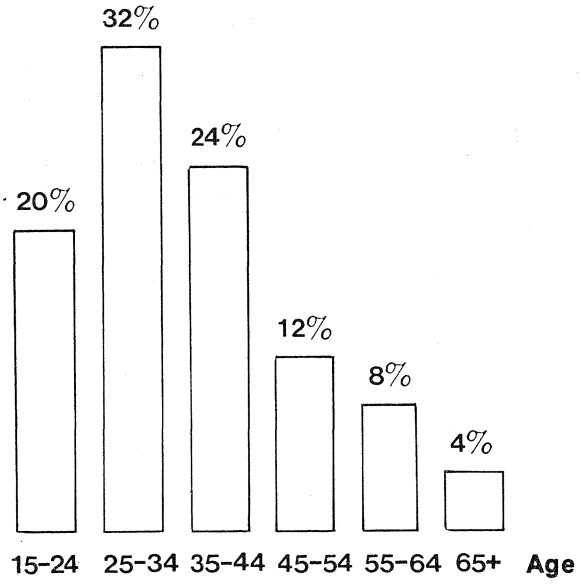
si **Sx** = *F et*

si $u = 1$, *alors* **Sat** = -2
si $2 \leq u \leq 3$, *alors* **Sat** = -1
si $4 \leq u \leq 5$, *alors* **Sat** = 0
si $6 \leq u \leq 8$, *alors* **Sat** = 1
si $9 \leq u \leq 12$, *alors* **Sat** = 2.

Une telle règle se traduit très facilement par un programme informatique; les ordinateurs peuvent choisir des nombres au hasard uniformément [2], effectuer des tests ou des comparaisons sur ces nombres, et se conformer aux résultats de ces tests pour imprimer les valeurs choisies des variables. À défaut de disposer d'un ordinateur, on peut aussi appliquer une telle règle manuellement à l'aide d'une table de nombres au hasard ou à l'aide d'une roulette de simulation.

Distribution des variables Ag, Scol, Rev et Opin.

En plus de simuler le sexe et le degré de satisfaction d'un répondant selon les règles données en exemple au paragraphe précédent, il faut simuler son âge, son niveau de scolarité, son revenu et son opinion. Puisque **Ag** est une variable indépendante de **Sx** et de **Sat**, on peut se donner une distribution arbitraire (mais plausible) pour cette variable dans l'ensemble de la population visée. Puisqu'il s'agit d'un sondage d'opinion, on peut supposer que tous les individus de la population visée sont âgés de 15 ans ou plus. On peut répartir les individus par groupes d'âge de la manière suivante:



À l'intérieur de chaque groupe d'âge on peut, en simplifiant la réalité, agir comme si la distribution était uniforme; on suppose alors que le groupe 65+ est 65-80. Dans ce cas, la règle de simulation pour la variable **Ag** se décrit ainsi:

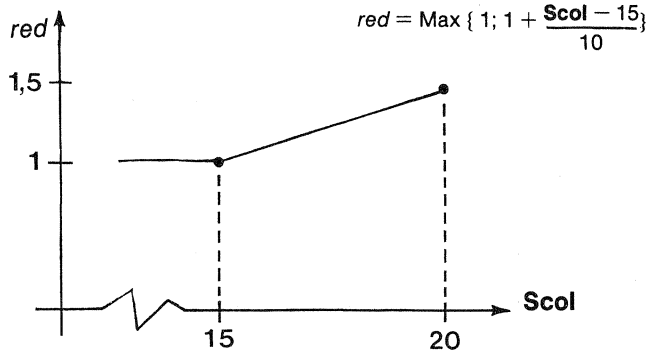
> *On choisit un nombre entier au hasard et uniformément entre 1 et 100. Soit N ce nombre. On choisit aussi, indépendamment, au hasard et uniformément, un nombre rationnel entre 0,00 et 1,00 (avec deux décimales après la virgule); soit u ce nombre.*
>
> *Si N ≤ 20, alors* **Ag** *= 15 + 9u, arrondi à l'entier le plus près;*
> *si 20 < N ≤ 52, alors* **Ag** *= 25 + 9u, arrondi à l'entier le plus près;*
> *si 52 < N ≤ 76, alors* **Ag** *= 35 + 9u, arrondi à l'entier le plus près;*
> *si 76 < N ≤ 88, alors* **Ag** *= 45 + 9u;*
> *si 88 < N ≤ 96, alors* **Ag** *= 55 + 9u;*
> *si 96 < N, alors* **Ag** *= 65 + 15u.*

Après avoir ainsi obtenu l'âge d'un répondant, on détermine son degré de scolarité. Ici, le modèle doit être suffisamment raffiné pour qu'un répondant âgé de 18 ans n'ait pas 20 ans de scolarité! En fait, dans le modèle, la scolarité d'un répondant s'obtient en retranchant 5 ans à son âge (l'école commence à 5 ans), et en retranchant un autre nombre qui varie aléatoirement suivant une distribution de Poisson [3]; dans tous les cas, la scolarité est d'au plus 20 ans; voici la règle de simulation pour **Scol**:

Si **Ag** *≤ 16, alors* **Scol** *=* **Ag** *− 5 −* Poiss($\lambda = 1$)
si **Ag** *= 17 ou 18, alors* **Scol** *=* **Ag** *− 5 −* Poiss($\lambda = 1{,}6$)
si 19 *≤* **Ag** *≤ 25, alors* **Scol** *=* **Ag** *− 5 −* Poiss($\lambda = 3$)
si 25 < **Ag**, *alors* **Scol** *= Min.{20;* Poiss($\lambda = 10$)*}*.

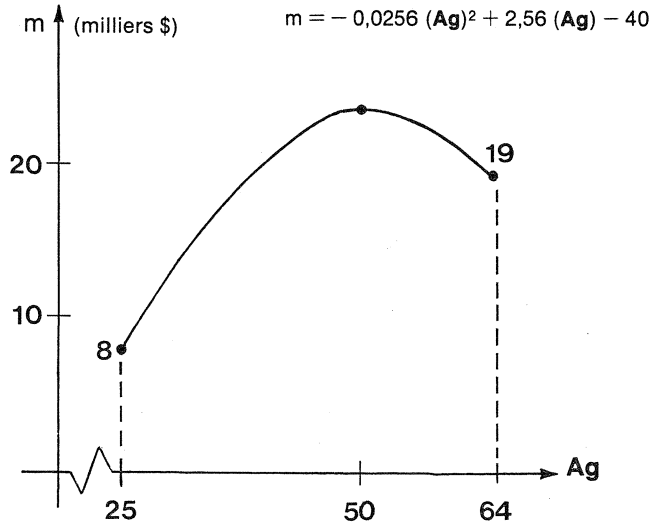
Les paramètres λ ont été choisis empiriquement de manière à rendre la distribution de la variable **Scol** plausible.

La variable **Rev** dépend à la fois de la variable **Ag** et de la variable **Scol**. On commence par simuler le revenu d'un répondant uniquement en fonction de son *âge* (et de facteurs aléatoires) et ensuite on *redresse* ce revenu en le multipliant par un coefficient (*red*) qui varie entre 1 et 1,5 uniquement selon la *scolarité* du répondant; cependant, le redressement n'est effectué que pour les répondants âgés de plus de 24 ans. Voici le modèle adopté pour le coefficient *red*:



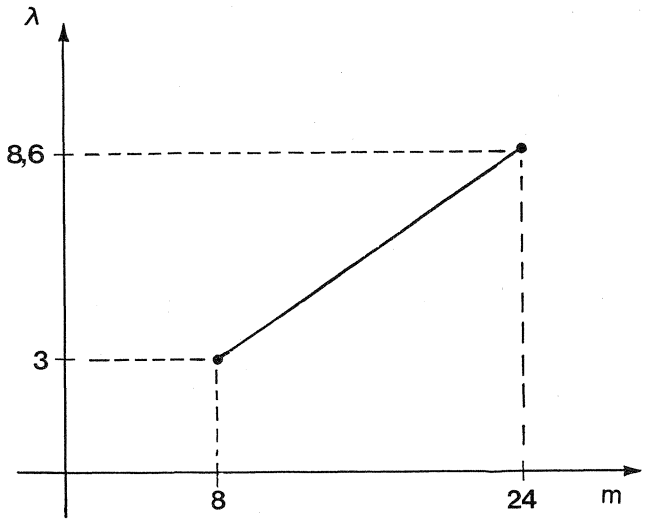
coefficient de redressement du revenu selon la scolarité: *red*

Pour simuler le revenu en fonction de l'âge, on choisit d'avance de faire varier le *revenu moyen* m à chaque âge selon la parabole suivante, pour les répondants âgés de moins de 65 ans:



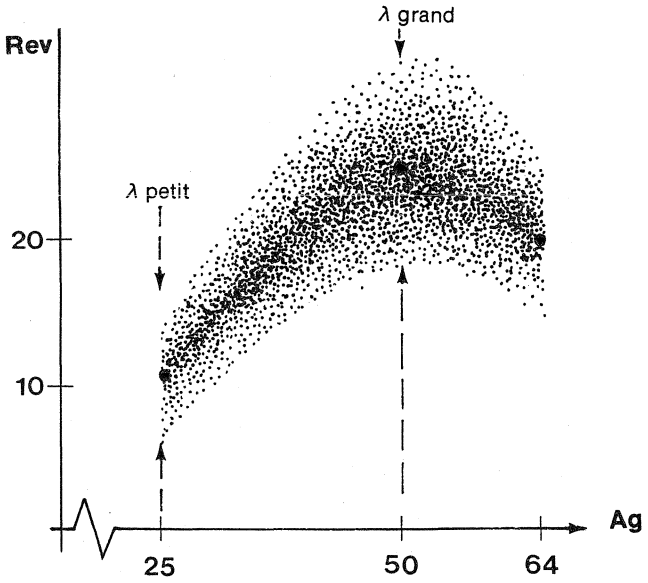
revenu moyen (avant redressement) à chaque âge

La dispersion des revenus à chaque âge autour de la moyenne m peut varier ainsi en fonction de m:



λ mesure la dispersion des revenus autour de la moyenne m.

De cette façon, on espère que le diagramme de dispersion des variables **Rev** et **Ag** aura l'allure suivante dans l'ensemble de la population visée par le sondage:



revenu avant redressement en fonction de l'âge pour les répondants âgés de 25 à 64 ans.

Pour les répondants âgés de 65 ans ou plus, on calcule un revenu minimal de 5 (en milliers de $) auquel on ajoute un revenu aléatoire qui décroît selon l'âge par un facteur noté *aba*. Pour les répondants âgés de moins de 25 ans, on calcule simplement une bourse d'au moins 2 (milliers de $) à laquelle s'ajoute un revenu de travail aléatoire distribué suivant une loi triangulaire. Voici le détail de la règle de simulation pour la variable **Rev**.

Si **Ag** ≤ 24, *alors* **Rev** $= \frac{v_1 + v_2}{2}$ *(valeur entière),*

où v_1 *et* v_2 *sont deux nombres aléatoires choisis uniformément et indépendamment*[4] *entre les bornes 2 et* $(2/3)$**Ag** $- 2$.

Si $25 \leq$ **Ag** ≤ 64,
alors **Rev** $= red(m - 2\lambda + 2$ Poiss$(\Lambda))$ *(valeur entière),*
où $red = Max\ \{1;\ 1 + \frac{\textbf{Scol} - 15}{10}\}$

$$m = -0{,}0256\ (\textbf{Ag}^2) + 2{,}56\ \textbf{Ag} - 40$$

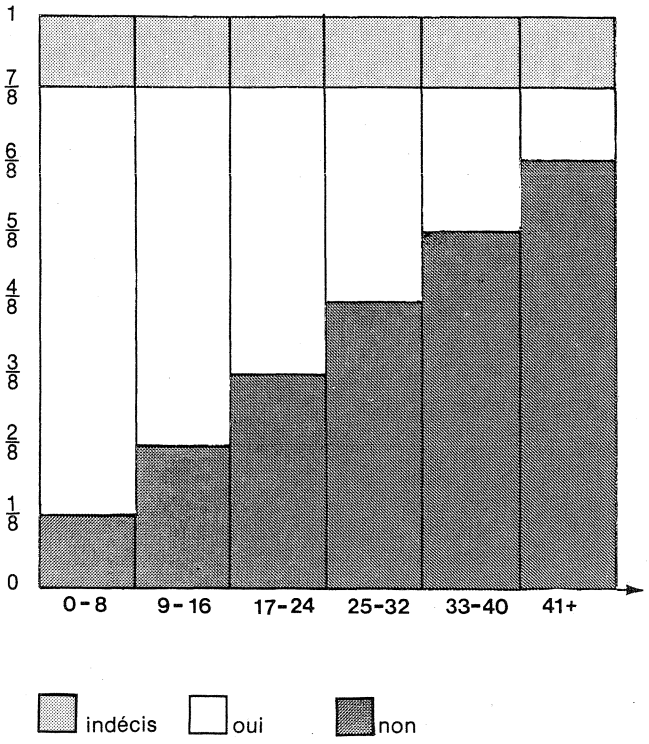
$$\lambda = 0{,}3636m + 0{,}0909$$

Si $65 \leq$ **Ag**,
alors **Rev** $= 5 + (res)(aba)\ 2$ Poiss$(\lambda = 7)$
$red = Max\ \{1;1 + \frac{\textbf{Scol} - 15}{10}\}$

$$aba = -0{,}0225\ (\textbf{Ag}) + 2{,}4625$$

(n.b.: la plupart des coefficients ont été choisis empiriquement de manière à rendre le modèle plausible).

Enfin, il ne reste à décrire que la règle de simulation de la variable **Opin**, qui dépend de la variable **Rev**. On a choisi (arbitrairement) la dépendance suivante entre **Opin** et **Rev**:



Ainsi, selon ce diagramme, la proportion des *indécis* est la même (1/8) dans toutes les catégories de revenus mais la proportion des *non* augmente avec le revenu.

Voici le détail de la règle de simulation pour la variable **Opin**, dictée par le diagramme donné ci-dessus:

On choisit un nombre entier au hasard, uniformément, entre 1 et 8; soit N ce nombre;
si **Rev** ≤ 8 *et si* $N = 1$, *alors* **Opin** = *non;*
si **Rev** ≤ 8 *et si* $N = 8$, *alors* **Opin** = *ind.;*
si **Rev** ≤ 8 *et si* $2 \leq N \leq 7$, *alors* **Opin** = *oui;*
si $9 \leq$ **Rev** ≤ 16 *et si* $N \leq 2$, *alors* **Opin** = *non;*
si $9 \leq$ **Rev** ≤ 16 *et si* $N = 8$, *alors* **Opin** = *ind.;*
si $9 \leq$ **Rev** ≤ 16 *et si* $3 \leq N \leq 7$, *alors* **Opin** = *oui;*
etc... jusqu'à:
si $41 \leq$ **Rev** *et si* $N \leq 6$, *alors* **Opin** = *non;*
si $41 \leq$ **Rev** *et si* $N = 8$, *alors* **Opin** = *ind.;*
si $41 \leq$ **Rev** *et si* $N = 7$, *alors* **Opin** = *oui.*

No	Sexe (Sx)	Âge (Ag)	Scolarité (Scol)	Revenu (Rev) milliers $	Satisfaction (Sat)	Opinion A = oui B = non C = indécis
157	M	50	12	18	0	B
158	F	29	14	19	-2	A
159	M	30	10	8	-1	A
160	M	29	8	9	-1	C
161	M	30	9	12	1	B
162	F	48	10	22	-1	B
163	F	30	13	10	1	B
164	M	31	9	10	-2	B
165	M	38	12	17	-2	A
166	F	32	8	10	-1	C
167	F	37	10	23	1	B
168	M	23	16	9	1	A
169	F	65	7	33	1	A
170	M	22	15	6	-1	A
171	M	52	13	28	-1	B
172	F	65	3	17	1	A
173	F	26	14	10	2	C
174	F	23	14	13	2	A
175	F	29	7	11	0	B
176	M	49	7	18	1	A
177	M	69	14	12	-2	B
178	M	50	12	24	-2	A
179	M	26	11	14	-2	A
180	M	33	8	18	2	B
181	F	61	12	16	1	B
182	F	52	10	20	2	B
183	M	46	3	16	-1	B
				10	1	B
					-2	A

Résultats de la simulation

Le tableau de la page précédente fournit les résultats de la simulation, tels qu'imprimés par un ordinateur qui a suivi les règles énoncées précédemment pour les six variables du modèle. Seule la règle de simulation de la variable âge (**Ag**) a été légèrement modifiée; de plus les différentes colonnes représentent des groupements naturels de valeurs qui facilitent la lecture lors de l'analyse des données (tableaux de contingence, calculs de moyennes pour des sous-catégories de répondants, représentations graphiques, etc.). La concordance entre les distributions théoriques et les distributions observées lors de la simulation de 2 000 répondants est excellente. On ne donne dans ce tableau qu'un extrait des résultats.

[1] Ces données font l'objet de plusieurs travaux pratiques pour les étudiants de niveau collégial et ont été publiées dans *Ateliers de probabilités et statistiques*, collection Mathécrit, Modulo éditeur, août 1981.

[2] En fait, ce sont des nombres *pseudo-aléatoires.*

[3] Les ordinateurs peuvent simuler des variables de Poisson; autrement, on peut utiliser les tables de Poisson que l'on trouve dans la plupart des manuels de statistiques.

[4] $\frac{U_1 + U_2}{2}$ suit alors une distribution de forme triangulaire. ■