Cahiers québécois de démographie

La nature et l'interprétation des variables indépendantes fonction du temps en démographie
The nature and meaning of time-varying covariates in demography

Benoît Laplante

Volume 38, numéro 1, printemps 2009

Enjeux de l'analyse démographique et nouvelles pistes méthodologiques

URI : https://id.erudit.org/iderudit/039990ar
DOI : https://doi.org/10.7202/039990ar

Aller au sommaire du numéro

Éditeur(s)

Association des démographes du Québec

ISSN

0380-1721 (imprimé)
1705-1495 (numérique)

Découvrir la revue

Citer cet article

Laplante, B. (2009). La nature et l'interprétation des variables indépendantes fonction du temps en démographie. *Cahiers québécois de démographie*, *38*(1), 105–143. https://doi.org/10.7202/039990ar

Résumé de l'article

On lit souvent que les modèles de risque « tiennent compte du passé » ou « possèdent une mémoire ». Contrairement à une croyance répandue, ceci n'est vrai que pour le quotient instantané de base et non pour la variation du quotient en fonction d'une variable indépendante fonction du temps (VIFT). Le quotient instantané de base varie en fonction du temps mesuré depuis l'origine, mais l'effet de la VIFT est markovien par construction : l'estimation de l'effet de la VIFT ne tient compte que de l'état occupé à chaque instant et pas des états occupés auparavant. En conséquence, son effet n'est pas conditionnel aux états occupés antérieurement. L'article examine cette question en adoptant le point de vue de la statistique mathématique, notamment en distinguant la population *théorique* construite à l'aide de modèles de risque et la population *réelle* et *finie* dont sont tirés les échantillons de personnes. L'examen se fait au moyen d'un exemple simple : l'effet de la situation conjugale sur la première naissance étudié à partir des données de l'*Enquête sociale générale* de 2006.

Cahiers québécois de démographie
Vol. 38, n° 1, Printemps 2009, p. 105-143

La nature et l'interprétation des variables indépendantes fonction du temps en démographie

BENOÎT LAPLANTE*

On lit souvent que les modèles de risque « tiennent compte du passé » ou « possèdent une mémoire ». Contrairement à une croyance répandue, ceci n'est vrai que pour le quotient instantané de base et non pour la variation du quotient en fonction d'une variable indépendante fonction du temps (VIFT). Le quotient instantané de base varie en fonction du temps mesuré depuis l'origine, mais l'effet de la VIFT est markovien par construction : l'estimation de l'effet de la VIFT ne tient compte que de l'état occupé à chaque instant et pas des états occupés auparavant. En conséquence, son effet n'est pas conditionnel aux états occupés antérieurement. L'article examine cette question en adoptant le point de vue de la statistique mathématique, notamment en distinguant la population *théorique* construite à l'aide de modèles de risque et la population *réelle* et *finie* dont sont tirés les échantillons de personnes. L'examen se fait au moyen d'un exemple simple : l'effet de la situation conjugale sur la première naissance étudié à partir des données de l'*Enquête sociale générale* de 2006.

English abstract, p. 143

On entend par variable indépendante fonction du temps (VIFT), l'utilisation, comme variable indépendante dans un modèle statistique, d'un caractère dont la modalité peut changer pendant que l'individu est considéré à risque de vivre l'événement qu'on étudie. L'usage des VIFT s'est répandu en démographie avec l'analyse des biographies et plus particulièrement avec le modèle de Cox (Cox, 1972) qui, le premier, a permis de les utiliser avec le sens qu'on leur donne couramment aujourd'hui. La VIFT est peut-être l'élément le plus riche de l'analyse des biographies puisqu'elle permet d'estimer l'effet des différentes modalités d'un caractère sur la survenue d'un événement — par exemple l'effet de la situation conjugale des femmes sur le risque d'avoir le premier enfant — en tenant compte du fait que le temps passé à risque par un individu peut-être réparti entre plusieurs des modalités de ce caractère : la portion de la biographie qui précède la naissance du premier enfant est habituellement

* Laboratoire d'études de la population, Centre-Urbanisation, culture et société, Institut national de la recherche scientifique, Montréal, Québec, Canada.

constituée de combinaison d'un ou plusieurs épisodes de vie solitaire, de vie en union de fait ou de mariage.

En dernière analyse, l'effet de la VIFT s'interprète exactement comme celui d'une variable indépendante ordinaire, mais cette similitude cache des particularités qui gagnent à être explicitées. Ainsi, on lit souvent que les modèles de risque « tiennent compte du passé » ou « possèdent une mémoire ». Ceci est vrai pour le quotient instantané[1] de base, par exemple la variation de ce quotient en fonction de l'âge. Ce n'est pas vrai pour la variation du quotient instantané en fonction d'une VIFT. Le fait qu'on possède, dans les données, de l'information sur les différentes modalités de la VIFT que chaque individu a occupées *successivement* peut amener à croire que le calcul du coefficient associé à chacune des modalités de la VIFT tient compte de la trajectoire de chaque individu. En réalité, le calcul du coefficient associé à une VIFT tient compte du temps passé à risque par chaque individu dans chacune des modalités de la VIFT, mais ne tient compte ni de l'ordre dans lequel ces modalités ont été occupées, ni du fait qu'elles aient été occupées successivement ou pas.

Dissiper cette confusion exige un examen raisonné et pas une simple assertion. La VIFT a été introduite en démographie avec des modèles statistiques qui servent à étudier les événements que la démographie étudie plus naturellement avec la table d'extinction. De ce fait, la VIFT ne se comprend vraiment que lorsqu'on la considère comme un modèle statistique qui repose sur la distinction et le rapport entre *population théorique* et *population finie*. Le propos de cet article est de faire comprendre la nature de la VIFT en illustrant cette idée à partir d'un exemple simple, l'âge à la naissance du premier enfant tel qu'on peut l'étudier à partir des données d'une enquête biographique rétrospective.

Dans la première partie de l'article, nous abordons la table d'extinction du point de vue de la statistique mathématique, ce qui nous permet de l'interpréter au moyen des notions de population théorique, de population réelle et d'échantillon, puis de l'envisager comme la réalisation d'une variable aléatoire. Dans la seconde partie, nous abordons la VIFT dans la table d'extinction avant d'examiner comment elle apparaît dans le modèle de Cox et dans les modèles paramétriques. En conclusion, nous présentons brièvement deux manières d'utiliser la VIFT afin d'intégrer, aux équations des modèles statistiques, des représentations plus complexes des trajectoires qui constituent les biographies.

1. Voir l'annexe pour le sens de cette expression.

LA TABLE D'EXTINCTION COMME MODÈLE STATISTIQUE

La table d'extinction est utilisée depuis le 17e siècle et, dans l'approche classique, elle n'est pas pensée comme un modèle statistique, mais plutôt comme le résultat d'un arrangement convenable des données (Pressat, 1973 : 17) ; ses propriétés statistiques ne sont étudiées que depuis peu (Lawless, 1982 : 53). Ce que nous proposons dans cette section n'est pas une étude formelle des propriétés statistiques de la table d'extinction, mais plutôt un exercice, qu'on qualifiera d'heuristique ou de didactique, dont le but est de la présenter du point de vue de la statistique mathématique. Cette étape est un peu longue, mais elle est nécessaire à la suivante, puisqu'il n'est pas possible d'élucider ce que sont les VIFT sans d'abord envisager la table d'extinction elle-même sous l'angle de la statistique mathématique.

La démarche est simple : on examine un caractère — dans notre exemple, l'âge des femmes canadiennes à la naissance de leur premier enfant — avec l'outil le plus ordinaire qu'on puisse imaginer pour examiner un caractère quantitatif — la distribution des effectifs par classes d'âge — et, en raisonnant sur les imperfections de cette distribution telles qu'on se les représente en statistique mathématique, on transforme la distribution de fréquences en table d'extinction. La table qui apparaît à la fin de la démarche n'a pas la forme à laquelle les démographes sont habitués, mais elle y apparaît clairement comme la réalisation d'une variable aléatoire.

Population théorique, population réelle et échantillon

Les quatre premières colonnes du tableau 1 — qui se détachent par leur trame — donnent les informations qu'on utilise couramment pour décrire la distribution d'un caractère quantitatif dans un échantillon[2]. Il s'agit ici de l'âge à la première naissance chez les 8 472 femmes de l'échantillon de l'*Enquête sociale générale* de 2006 (ESG dans la suite du texte) dont on sait qu'elles ont eu leur premier enfant entre 15 et 40 ans. De la première à la quatrième colonne, on trouve les limites de chaque classe d'âge, la distribution des effectifs, la distribution des fréquences et la distribution des fréquences cumulées.

Les fréquences et les fréquences cumulées sont des proportions. La distribution des fréquences donne la proportion des individus qui change

2. Voir l'annexe pour le sens du mot « caractère ».

TABLEAU 1 L'âge à la naissance du premier enfant – Femmes âgées de 15 à 80 ans au moment de l'enquête

	Distribution tronquée à gauche et à droite						Distribution tronquée à gauche et censurée à droite					
T	n_t	f_t	F_t	S_t	h_t	H_t	n_t	f_t	F_t	S_t	h_t	H_t
De 15 ans à moins de 16	47	0,0056	0,0056	0,9944	0,0056	0,0056	193	0,0154	0,0154	0,9846	0,0154	0,0154
De 16 ans à moins de 17	142	0,0168	0,0223	0,9777	0,0169	0,0224	292	0,0233	0,0388	0,9612	0,0237	0,0391
De 17 ans à moins de 18	303	0,0358	0,0581	0,9419	0,0366	0,059	436	0,0349	0,0736	0,9264	0,0363	0,0754
De 18 ans à moins de 19	426	0,0503	0,1084	0,8916	0,0534	0,1124	549	0,0439	0,1175	0,8825	0,0474	0,1228
De 19 ans à moins de 20	590	0,0697	0,1781	0,8219	0,0781	0,1906	701	0,0561	0,1736	0,8264	0,0635	0,1863
De 20 ans à moins de 21	705	0,0833	0,2613	0,7387	0,1013	0,2919	832	0,0665	0,2401	0,7599	0,0805	0,2668
De 21 ans à moins de 22	619	0,0731	0,3344	0,6656	0,099	0,3908	748	0,0598	0,2999	0,7001	0,0787	0,3455
De 22 ans à moins de 23	685	0,0809	0,4153	0,5847	0,1215	0,5124	797	0,0637	0,3637	0,6363	0,091	0,4366
De 23 ans à moins de 24	649	0,0766	0,492	0,508	0,1311	0,6434	756	0,0605	0,4241	0,5759	0,095	0,5316
De 24 ans à moins de 25	582	0,0687	0,5607	0,4393	0,1353	0,7787	692	0,0553	0,4794	0,5206	0,0961	0,6277
De 25 ans à moins de 26	660	0,0779	0,6386	0,3614	0,1774	0,9562	801	0,064	0,5435	0,4565	0,123	0,7507
De 26 ans à moins de 27	489	0,0577	0,6964	0,3036	0,1598	1,116	597	0,0477	0,5912	0,4088	0,1046	0,8553
De 27 ans à moins de 28	516	0,0609	0,7573	0,2427	0,2007	1,3167	610	0,0488	0,64	0,36	0,1193	0,9746
De 28 ans à moins de 29	395	0,0466	0,804	0,196	0,1922	1,5089	514	0,0411	0,6811	0,3189	0,1142	1,0888
De 29 ans à moins de 30	379	0,0448	0,8487	0,1513	0,2283	1,7372	465	0,0372	0,7183	0,2817	0,1166	1,2054
De 30 ans à moins de 31	296	0,035	0,8837	0,1163	0,2311	1,9683	378	0,0302	0,7485	0,2515	0,1073	1,3127
De 31 ans à moins de 32	245	0,0289	0,9126	0,0874	0,2487	2,217	317	0,0253	0,7739	0,2261	0,1008	1,4135
De 32 ans à moins de 33	207	0,0244	0,9371	0,0629	0,2797	2,4967	290	0,0232	0,7971	0,2029	0,1025	1,516
De 33 ans à moins de 34	157	0,0185	0,9556	0,0444	0,2946	2,7913	207	0,0166	0,8136	0,1864	0,0816	1,5976
De 34 ans à moins de 35	113	0,0133	0,9689	0,0311	0,3005	3,0918	167	0,0134	0,827	0,173	0,0716	1,6692
De 35 ans à moins de 36	84	0,0099	0,9789	0,0211	0,3194	3,4112	145	0,0116	0,8386	0,1614	0,067	1,7362
De 36 ans à moins de 37	73	0,0086	0,9875	0,0125	0,4078	3,819	136	0,0109	0,8494	0,1506	0,0674	1,8036
De 37 ans à moins de 38	53	0,0063	0,9937	0,0063	0,5	4,319	113	0,009	0,8585	0,1415	0,06	1,8636
De 38 ans à moins de 39	30	0,0035	0,9973	0,0027	0,566	4,8851	87	0,007	0,8654	0,1346	0,0492	1,9127
De 39 ans à moins de 40	23	0,0027	1	0	1	5,8851	1683	0,1346	1	0	1	2,9127
Somme	8 468						12 506					

Source : Statistique Canada, Enquête sociale générale de 2006.

d'état *au cours* de chaque intervalle. Lorsqu'on s'intéresse à un caractère qui peut s'interpréter comme le temps passé dans l'état d'origine, la distribution des fréquences cumulées a un sens précis : elle donne la proportion des individus qui a quitté l'état d'origine *à la fin* de chaque intervalle.

Règle générale, on interprète les fréquences et les fréquences cumulées comme des proportions de l'échantillon et, si l'échantillon est probabiliste, on peut les interpréter comme des proportions de la population dont l'échantillon a été tiré. Cette interprétation est simple à première vue, mais elle cache une difficulté importante. En statistique mathématique, on distingue deux types de population bien différents : les populations finies et les populations théoriques ou infinies. Par population finie, on entend habituellement une population réelle, par exemple la population du Canada telle qu'elle existe au moment où un échantillon en est tiré. Une population théorique est plutôt un être statistique défini par une loi de probabilité, dont on peut, en principe, tirer aussi bien une ou plusieurs populations finies qu'un ou plusieurs échantillons. Lorsqu'on utilise les valeurs d'un caractère recueillies au moyen d'un échantillon pour décrire la distribution de ce caractère dans la population — par exemple, l'âge des femmes canadiennes à la naissance de leur premier enfant à partir des données de l'ESG — on postule que l'échantillon a été tiré d'une population finie. On suppose alors que la distribution du caractère dans l'échantillon donne une image assez fidèle de la distribution de ce caractère dans la population et que la seule source d'imprécision est l'erreur d'échantillonnage. On ne raisonne pas de manière différente lorsqu'on construit une distribution à deux caractères, par exemple un tableau croisé.

On se trouve cependant à raisonner de manière tout à fait différente lorsqu'on utilise les informations recueillies auprès du même échantillon pour relier la distribution d'un caractère aux valeurs de certains autres caractères au moyen d'un modèle statistique. Dans ce cas, la distribution du caractère *dans la population réelle* n'est plus simplement vue comme une caractéristique de celle-ci qu'on peut décrire de manière raisonnable à partir de sa distribution dans l'échantillon. On voit plutôt le caractère comme une variable aléatoire, c'est-à-dire une variable dont chacune des valeurs est associée à une probabilité, et la relation entre cette variable aléatoire et les autres caractères comme une fonction mathématique composée d'un ensemble de relations systématiques et d'un mécanisme aléatoire régi par une loi de probabilité. En plus, on considère que la distribution du caractère *dans la population réelle* est une réalisation de cette variable aléatoire, c'est-à-dire le produit d'un tirage dans la distribution théorique de cette variable aléatoire telle qu'elle résulte du modèle. Autrement dit, on

considère que la population réelle et finie est elle-même un échantillon tiré dans une population théorique et infinie, et que l'échantillon au sens habituel est un échantillon tiré de cet échantillon. Dans ce contexte, on réserve le mot « population » employé absolument pour désigner la population réelle finie et on nomme « superpopulation » la population ou distribution théorique qui correspond au modèle (Korn et Graubard, 1999 : 62, 89-94 ; Binder et Roberts, 2003 : 31-34). La distribution du caractère dans l'échantillon au sens habituel devient alors, d'abord et avant tout, une approximation de la distribution théorique qui correspond au modèle et donc une approximation de la distribution de la variable aléatoire[3].

Par définition, une variable aléatoire est régie par une loi de probabilité. Une loi de probabilité associe une probabilité à chacune des valeurs d'une variable. On exprime le plus souvent une loi de probabilité par sa *fonction de densité* de la probabilité — généralement notée $f(t)$ — qui correspond à la distribution des fréquences. Toute loi de probabilité a également une *fonction de répartition* de la probabilité — généralement notée $F(t)$ — qui correspond à la distribution des fréquences cumulées. Toute loi de probabilité a également une *fonction de survie* ou *fonction de séjour* $S(t)$ ainsi qu'une *fonction de risque* $h(t)$ et une *fonction de risque cumulé* $H(t)$. La fonction de survie est le complément arithmétique de la fonction de répartition, c'est-à-dire $S(t) = 1 - F(t)$: elle donne la proportion de la population qui n'a pas quitté l'état d'origine à la fin de chaque intervalle. La fonction de risque donne l'intensité du processus qui régit le changement d'état au cours de chaque intervalle ; on peut l'obtenir à partir de plusieurs relations, mais on la définit habituellement comme le quotient[4] de la fonction de densité et de la fonction de survie, c'est-à-dire $h(t) = f(t) / S(t)$. La fonction de risque cumulée est l'intégrale du risque instantané, c'est-à-dire $H(t) = \int h(t)$; lorsque l'événement est renouvelable, on l'interprète comme le nombre moyen d'occurrences *à la fin* de chaque intervalle. Dans le tableau 1, on a calculé les différentes fonctions en appliquant les définitions générales.

3. Nous prenons ici le contrepied de Hoem (1985). La position de Hoem était discutable au moment où il a écrit son article ; elle est intenable depuis les travaux de Binder (1983, 1992) et de Rao et Wu (1988).
4. Le mot « quotient » est entendu ici au sens général qu'il a en mathématiques où il désigne le résultat d'une division. Dans la suite du texte, on utilisera le mot « quotient » surtout au sens spécial qu'il a en démographie où il désigne la probabilité de quitter l'état d'origine au cours d'un intervalle fini.

La loi de probabilité que suit une variable aléatoire peut appartenir à une famille de lois de probabilité paramétriques, par exemple la loi normale, la loi exponentielle, la loi de Weibull, etc. Dans ce cas, la fonction de densité, la fonction de répartition, la fonction de survie, la fonction de risque et la fonction de risque cumulé se représentent sous forme algébrique. On peut également admettre qu'une variable aléatoire suive une loi non paramétrique suffisamment décrite par la distribution de sa réalisation dans un échantillon. On considère alors que la loi de probabilité qui régit la variable aléatoire est suffisamment décrite par l'une ou l'autre des fonctions qu'on aura estimée à partir de l'échantillon. La distribution des fréquences sera alors un estimé de la fonction de densité, la distribution des fréquences cumulées sera un estimé de la fonction de répartition et ainsi de suite. C'est ainsi que s'interprètent les fonctions qu'on trouve dans les colonnes du tableau 1.

La table d'extinction comme objet statistique

Examinons maintenant les deux distributions du tableau 1. La première distribution est construite en ne retenant que les femmes qui ont eu leur premier enfant après l'âge de 15 ans et avant 40 ans, et chaque femme est classée à l'âge où elle a eu cet enfant. Comme nous l'avons déjà noté, il s'agit de la distribution de l'âge à la première naissance chez les femmes qui ont eu leur premier enfant entre 15 et 40 ans. La seconde distribution est construite en retenant toutes les femmes qui n'ont pas eu leur premier enfant avant l'âge de 15 ans et en classant chaque femme soit à l'âge où elle a eu son premier enfant, soit à l'âge le plus élevé où elle a été observée sans avoir eu son premier enfant.

Il est évident qu'aucune des deux distributions du tableau 1 n'est une bonne approximation de la distribution de la variable aléatoire dont on présume qu'a été tiré l'âge des femmes canadiennes à la naissance de leur premier enfant. La première ne donne que l'âge à la naissance du premier enfant des femmes qui ont eu cet enfant entre 15 et 40 ans, et on ne voit pas d'interprétation raisonnable de la seconde. Malgré cela, ces distributions sont importantes parce qu'elles illustrent les problèmes qui se posent lorsqu'on tente de décrire la distribution d'une variable aléatoire, comme l'âge à un événement démographique, à partir de la distribution de fréquences d'un échantillon biographique rétrospectif. Ces problèmes ont une solution évidente pour toute personne qui sait ce qu'est une table d'extinction, mais l'important, ici, n'est pas d'arriver rapidement à la solution, mais de comprendre comment le problème est perçu et nommé en statistique mathématique.

Dans son *Dictionnaire de statistique*, Morice (1968 : 31) définit comme suit la censure et la troncation, les deux concepts qu'utilisent les statisticiens pour décrire les distributions du tableau 1 :

> **Distribution tronquée.** Distribution obtenue à partir d'une distribution donnée en ignorant les parties situées en deçà (ou au-delà) d'une valeur donnée *a* (point de troncation). Certains auteurs distinguent *distribution tronquée* [on connaît seulement la distribution en deçà (ou au-delà) de la valeur a] et *distribution censurée* (on connaît de plus le nombre mais non la distribution dans la partie abandonnée).

La définition de Morice résume de la manière la plus simple et la plus claire qui soit les travaux de Fisher (1931) et de Hald (1949) qui ont introduit en statistique mathématique les notions de distribution tronquée et de distribution censurée ; on retrouve les mêmes définitions dans la version anglaise du *Dictionnaire de démographie* de R. Pressat (1985 : 223)[5].

Examinons à nouveau les distributions du tableau 1, mais cette fois-ci à la lumière des définitions de la troncation et de la censure. Au sens de ces définitions, il est tout d'abord évident que les deux distributions du tableau 1 sont tronquées à gauche : on exclut de chacune les premières naissances survenues avant l'âge de 15 ans. Examinons plus particulièrement la première distribution. On voit qu'en plus d'être tronquée à gauche, cette distribution est tronquée à droite, puisqu'en plus d'exclure les naissances survenues avant l'âge de 15 ans, elle exclut également les naissances survenues après 40 ans. Faisons de même pour la deuxième distribution. On voit qu'en plus d'être tronquée à gauche, cette distribution est censurée à droite : les femmes qui n'ont pas eu leur premier enfant avant 40 ans ne sont pas exclues, mais sont classées à la limite inférieure de la portion de la distribution où se trouvera l'âge auquel elles donneront naissance à leur premier enfant si elles le font. Puisque l'échantillon est composé de femmes âgées de 15 à 80 ans au moment de l'enquête et que les femmes nullipares sont comptées dans la classe d'âge à laquelle elles appartenaient au moment de l'enquête, la distribution a autant de points de censure qu'elle a de classes d'âge.

---

5. Il faut éviter de confondre ces deux notions avec l'« effet de troncature » dont traite Pressat (1979 : 63) qui est en fait un cas particulier de la censure (*cf.* Pressat, 1985 : 224). L'original du *Dictionnaire de démographie* ne traite pas de ces notions, comme il ne traite d'ailleurs d'aucun sujet de statistique mathématique. On trouve un traitement complet de la censure et de la troncation en analyse longitudinale dans Klein et Moeschberger (2003).

TABLEAU 2 La table d'extinction comme modèle statistique – L'âge à la naissance du premier enfant – Femmes âgées de 15 à 80 ans au moment de l'enquête

T	À risque au début	Ajouts	Retraits	Événements	r_t	m_t	q_t	$f(t)$	$F(t)$	$S(t)$	$H(t)$
De 15 ans à moins de 16	11 038	0	144	43	10 950,70	0,0039	0,0039	0,0039	0,0039	0,9961	0,0039
De 16 ans à moins de 17	10 851	0	150	132	10 724,90	0,0123	0,0122	0,0122	0,0161	0,9839	0,0162
De 17 ans à moins de 18	10 569	217	132	286	10 491,60	0,0273	0,027	0,0266	0,0426	0,9574	0,0435
De 18 ans à moins de 19	10 368	0	122	406	10 140,60	0,04	0,0393	0,0377	0,0803	0,9197	0,0835
De 19 ans à moins de 20	9 840	0	111	558	9 523,85	0,0586	0,057	0,0525	0,1328	0,8672	0,1421
De 20 ans à moins de 21	9 171	0	127	653	8 825,55	0,074	0,0716	0,0621	0,1949	0,8051	0,2161
De 21 ans à moins de 22	8 391	0	129	568	8 075,95	0,0703	0,0681	0,0549	0,2498	0,7502	0,2864
De 22 ans à moins de 23	7 694	281	111	614	7 518,45	0,0817	0,0788	0,0591	0,3089	0,6911	0,3681
De 23 ans à moins de 24	7 250	0	107	597	6 940,75	0,086	0,0829	0,0573	0,3661	0,6339	0,4541
De 24 ans à moins de 25	6 546	0	110	541	6 250,80	0,0865	0,0833	0,0528	0,4189	0,5811	0,5406
De 25 ans à moins de 26	5 895	0	141	603	5 557,55	0,1085	0,1032	0,06	0,4789	0,5211	0,6491
De 26 ans à moins de 27	5 151	0	106	447	4 889,10	0,0914	0,0876	0,0457	0,5246	0,4754	0,7405
De 27 ans à moins de 28	4 598	174	93	473	4 439,05	0,1066	0,1019	0,0484	0,573	0,427	0,8471
De 28 ans à moins de 29	4 206	0	119	372	3 982,85	0,0934	0,0896	0,0383	0,6113	0,3887	0,9405
De 29 ans à moins de 30	3 715	0	86	354	3 516,40	0,1007	0,0963	0,0374	0,6487	0,3513	1,0412
De 30 ans à moins de 31	3 275	0	82	278	3 088,50	0,09	0,086	0,0302	0,6789	0,3211	1,1312
De 31 ans à moins de 32	2 915	0	70	228	2 777,70	0,0821	0,0791	0,0254	0,7043	0,2957	1,2133
De 32 ans à moins de 33	2 617	82	81	193	2 528,55	0,0763	0,0735	0,0217	0,7261	0,2739	1,2896
De 33 ans à moins de 34	2 425	0	50	149	2 327,00	0,064	0,062	0,017	0,7431	0,2569	1,3536
De 34 ans à moins de 35	2 226	0	54	109	2 146,30	0,0508	0,0496	0,0127	0,7558	0,2442	1,4044
De 35 ans à moins de 36	2 063	0	60	79	2 000,50	0,0395	0,0388	0,0095	0,7653	0,2347	1,4439
De 36 ans à moins de 37	1 924	0	62	70	1 863,50	0,0376	0,037	0,0087	0,774	0,226	1,4815
De 37 ans à moins de 38	1 792	39	60	52	1 756,20	0,0296	0,0292	0,0066	0,7806	0,2194	1,5111
De 38 ans à moins de 39	1 719	0	57	30	1 676,65	0,0179	0,0177	0,0039	0,7845	0,2155	1,529
De 39 ans à moins de 40	1 632	0	1 609	23	1 577,71	0,0146	0,0145	0,0031	0,7876	0,2124	1,5436
Somme				7 858							

Source: Statistique Canada, Enquête sociale générale de 2006.

Cet examen des deux distributions du tableau 1 montre qu'on ne peut pas estimer la distribution de la variable aléatoire à partir de la distribution des fréquences du caractère, tel qu'il est observé dans l'échantillon, tout simplement parce qu'on ne connaît pas l'âge à la naissance du premier enfant de chacune des femmes qui le composent. Ce problème est général. La solution consiste à estimer la distribution de la variable aléatoire en appliquant la logique de la table d'extinction.

L'estimation des fonctions de la table d'extinction vue comme la réalisation d'une variable aléatoire

À strictement parler, le risque est la probabilité de changer d'état au cours d'un intervalle ; en ce sens, le quotient de la table d'extinction est un risque. Toujours à strictement parler, la fonction de risque est la limite de cette probabilité, lorsque la longueur de l'intervalle tend vers zéro, pour un événement non renouvelable ; pour un événement renouvelable, on parle habituellement de la fonction d'intensité.

En épidémiologie, on emploie habituellement l'expression « risque instantané » pour désigner la limite du quotient d'un intervalle lorsque la longueur de celui-ci tend vers zéro[6]. Lorsque l'intervalle tend vers zéro, le quotient — ou risque — qui est une probabilité, et le taux, qui n'en est pas une, se confondent ; pour cette raison, le risque instantané possède des propriétés qui appartiennent autrement soit au quotient — ou au risque — (on peut l'utiliser pour former un produit : $S_t = \prod (1 - h_k), k = 1 \ldots t$), soit au taux (on peut en faire la somme : $H_t = \sum h_k, k = 1 \ldots t$).

La théorie qui traite des modèles statistiques utilisés en analyse des biographies ne raisonne que sur les intervalles infinitésimaux, même pour le cas discret, et ignore ou contourne les particularités du découpage en intervalles finis qui est à la base de la table d'extinction, notamment celle qui force à distinguer le quotient du taux. Il n'y a donc pas de définition stricte de la fonction de risque pour la table d'extinction ; plutôt que la fonction de risque à proprement parler, nous calculons donc les taux (notés m_t) et les quotients (notés q_t).

Normalement, le questionnaire biographique mesure la durée des séjours dans l'état d'origine avec une précision plus grande que la longueur des intervalles d'une table. Cette précision permet de construire la table à partir des taux dont les dénominateurs — le temps passé à risque par l'en-

6. Voir l'annexe pour le sens des mots « risque », « taux » et « quotient » et des expressions qui leur sont associées.

semble des individus à risque dans chaque intervalle — sont mesurés avec précision. On trouve le temps total passé à risque au cours de chaque intervalle dans la colonne r_t du tableau 2. Le taux est obtenu tout simplement en divisant le nombre des événements survenus au cours de l'intervalle par la quantité de temps passé à risque au cours de cet intervalle par l'ensemble des individus qui y ont été à risque[7]. On se sert directement des taux pour calculer la fonction de risque cumulé $H(t)$.

Connaître exactement le temps passé à risque par chaque individu permet également de calculer le quotient à partir du taux sans faire d'hypothèse sur la fraction moyenne de l'intervalle qui est passée à risque par les individus qui changent d'état au cours de cet intervalle. On peut ainsi obtenir le quotient en suivant Chiang (1984 : 72, 119), dont l'équation 2.4 se ramène à $q_t = m_t / [1 + (1 - a_t) \cdot m_t]$ lorsque l'intervalle est d'une unité et dans laquelle a_t est cette fraction que l'information disponible permet d'estimer ; on nomme parfois cette fraction le « coefficient de répartition », mais, à part le nom, elle n'a rien en commun avec la fonction de répartition. On calcule les fonctions de survie, de répartition et de densité à partir des quotients. Pour ne pas surcharger le tableau, nous n'avons pas inclus les valeurs de a_t.

---

7. Lorsqu'on construit une table d'extinction à partir de données agrégées, il est d'usage de calculer le taux en divisant le nombre des événements survenus au cours de l'intervalle par la taille de la population à risque au centre de l'intervalle. Ce dénominateur est en fait une approximation du temps passé à risque par la population à risque au cours de l'intervalle (Pressat, 1979 : 243). Le choix de ce dénominateur repose sur un postulat irréaliste : le nombre des événements serait réparti également de part et d'autre du centre de l'intervalle. On comprend ce problème en réfléchissant à la mortalité d'une cohorte en l'absence de migration. En cours d'intervalle, la population à risque de décéder diminue chaque jour puisqu'une partie de la cohorte décède chaque jour. Pour que le nombre des décès survenus dans une des moitiés de l'intervalle soit égal au nombre des décès survenus dans l'autre, il faudrait que le taux de mortalité soit plus élevé dans la seconde moitié que dans la première ; plus généralement, il faudrait en fait que le taux augmente de manière exponentielle tout au long de l'intervalle. Cette difficulté se contourne facilement en abordant la question comme un problème de croissance ou de décroissance. On fait alors de la taille de la population à risque à la fin de l'intervalle — en pratique, la taille de cette population au début de l'intervalle suivant — une fonction de la taille de cette population au début de l'intervalle et d'un taux que nous noterons x (c.-à-d. $P_{t+1} = P_t^x$) qu'on isole en utilisant une propriété des logarithmes qui permet d'écrire $\ln(P_t) = \ln(P_{t+1})$ puis $x = \ln(P_{t+1}) / \ln(P_t)$. Contrairement au taux obtenu en divisant le nombre des événements survenus au cours de l'intervalle par la taille de la population à risque au centre de l'intervalle, ce taux est équivalent au taux qu'on obtient en divisant le nombre des événements par le temps passé à risque au cours de l'intervalle : il est constant tout au long de l'intervalle.

Dans notre exemple, la troncation à gauche est un choix. Dans l'échantillon comme dans la population réelle, certaines femmes donnent naissance à leur premier enfant avant l'âge de 15 ans. Nous choisissons néanmoins de les exclure comme cela se fait couramment dans l'étude de la fécondité. En ne retenant que les femmes qui n'ont pas eu leur premier enfant avant l'âge de 15 ans, on estime donc volontairement une distribution tronquée à gauche ; autrement dit, on fixe l'origine de la table au point de troncation.

Le problème de la censure à droite se règle comme il se doit en retirant du groupe à risque, à l'âge qu'elles avaient au moment de l'enquête, les femmes qui, au moment de l'enquête, n'avaient toujours pas eu leur premier enfant. Ce sont les retraits du tableau 2.

TABLEAU 3 Âge au début de la vie au Canada – Femmes âgées de 15 à 80 ans au moment de l'enquête

Classe	Valeur	Effectifs
Né au Canada ou citoyen de naissance	0	10 511
Moins de 5 ans	2,5	244
De 5 ans à moins de 10	7,5	166
De 10 ans à moins de 15	12,5	158
De 15 ans à moins de 20	17,5	223
De 20 ans à moins de 25	22,5	357
De 25 ans à moins de 30	27,5	306
De 30 ans à moins de 35	32,5	228
De 35 ans à moins de 40	37,5	153
40 ans ou plus		164
Inconnu		47
Total		12 557

Source : Statistique Canada, Enquête sociale générale de 2006.

Nous étudions la naissance du premier enfant au Canada. Près de 16 % des femmes de l'échantillon de l'ESG âgées de 15 à 80 ans sont nées à l'étranger. On ne peut les considérer à risque d'avoir un enfant au Canada que si elles s'y sont établies avant l'âge de quarante ans et avant d'avoir donné naissance à leur premier enfant. Nous ne connaissons pas l'âge exact auquel les femmes nées à l'étranger se sont établies au Canada : celui-ci nous est fourni en classes de cinq ans. Le tableau 3 donne la distribution

de l'âge auquel les femmes nées à l'étranger se sont établies au Canada et également le nombre de celles qui y sont nées. On y trouve également le centre de chaque classe, que nous utilisons pour fixer l'âge à partir duquel les femmes nées à l'étranger peuvent donner naissance à leur premier enfant au Canada plutôt qu'à l'étranger. On doit tenir compte de ce fait dans le calcul des taux. Les femmes nées à l'étranger sont donc intégrées au groupe à risque au moment où elles arrivent au Canada, ce qui revient à dire qu'on les intègre à la table en appliquant la logique des entrées échelonnées. Ce sont les ajouts du tableau 2.

L'échantillon de l'ESG comprenait 12 557 femmes âgées de 15 à 80 ans au moment de l'enquête. Nous avons éliminé 47 femmes nées à l'étranger parce qu'on ne sait pas à quel âge elles se sont établies au Canada, 591 autres femmes nées à l'étranger parce qu'elles ont eu leur premier enfant avant de s'établir au Canada et 11 parce qu'on ne sait pas à quel âge elles ont eu leur premier enfant. Nous avons exclu à dessein les 41 femmes nées au Canada ou établies au Canada avant l'âge de 15 ans qui ont eu leur premier enfant avant cet âge. Enfin, nous avons ignoré les 36 femmes nées à l'étranger qui se sont établies au Canada alors qu'elles avaient au moins quarante ans et qu'elles n'avaient toujours pas eu leur premier enfant. Nous avons donc estimé les taux à partir de l'information recueillie auprès des 11 831 femmes de l'échantillon qui ont été à risque de donner naissance à leur premier enfant au Canada après leur quinzième anniversaire ; 7 858 de ces femmes ont donné naissance à leur premier enfant avant leur quarantième anniversaire.

Lorsqu'on construit une table d'extinction, on ne s'amuse habituellement pas à calculer la table de survie d'un événement renouvelable ou l'indice synthétique d'un événement non renouvelable. Lorsqu'on utilise un modèle statistique, les différentes fonctions de la variable aléatoire existent et peuvent intervenir dans l'estimation du modèle même si elles n'ont pas ce que les statisticiens nomment une interprétation naturelle. On peut ainsi étudier un événement renouvelable au moyen du modèle de Cox ou des modèles paramétriques conçus en principe pour l'étude des événements non renouvelables (p. ex. les modèles basés sur les lois exponentielles, de Weibull, de Gompertz, etc.) dont la vraisemblance est construite à partir du rapport entre la fonction de densité et la fonction de survie[8]. On peut également étudier un événement non renouvelable au moyen d'un modèle conçu en principe pour étudier les événements renouvelables,

8. Nous expliquons cette idée dans la section suivante.

comme le modèle de Poisson ; les logiciels d'analyse multiniveau comme *MLwiN* ou *gllamm* estiment les modèles de risque de cette façon.

La table construite à partir de données recueillies au cours d'une enquête biographique n'est pas la table d'une cohorte, puisqu'elle est établie à partir des informations recueillies auprès d'individus qui appartiennent à plusieurs cohortes. Elle n'est pas non plus une table du moment puisque les événements utilisés dans les numérateurs et le temps à risque utilisé dans les dénominateurs sont répartis entre les différentes classes d'âge traversées par les individus au cours de leur vie. Les individus qui appartiennent aux différentes cohortes sont observés rétrospectivement jusqu'à l'âge qu'ils ont au moment de l'enquête. La composition par cohortes de la population théorique décrite par la table change donc selon les classes d'âge : les taux des classes d'âge inférieures sont estimés en combinant l'expérience de toutes les cohortes alors que les taux des classes d'âge supérieures ne sont estimés qu'à partir de l'expérience des cohortes les plus âgées. Pour le meilleur et pour le pire, la variable aléatoire, ou la population théorique, que décrivent les fonctions résume l'expérience de plusieurs cohortes nées dans la société étudiée et celle des immigrants dans la mesure où elle est vécue dans la société étudiée.

LA VARIABLE INDÉPENDANTE FONCTION DU TEMPS

Nous disposons maintenant de tous les éléments nécessaires pour aborder le problème de la VIFT. La variable indépendante fonction du temps est un caractère dont la modalité peut changer pendant que l'individu est considéré à risque de vivre l'événement qu'on étudie. Dans notre exemple, les femmes sont considérées à risque de donner naissance à leur premier enfant dès l'âge de quinze ans. À cet âge, la plupart d'entre elles ne sont pas mariées et ne vivent pas en union de fait, mais la plupart auront été mariées ou auront vécu en union de fait avant d'avoir leur premier enfant. Ceci revient à dire que pour la plupart des femmes, le temps à risque de donner naissance au premier enfant sera réparti entre deux ou trois des modalités de la situation conjugale : vivre sans conjoint, être mariée et vivre avec son époux ou bien vivre en union de fait. Formellement, ce problème est analogue à celui que pose la migration interne dans l'étude de la mortalité et il se résout de la même manière : on construit une table d'extinction pour chacune des modalités du caractère comme on construit une table de mortalité pour chaque région, et on déplace l'individu d'une table à l'autre à l'âge où il passe d'une modalité du caractère à une autre, exactement comme on déplace un individu d'une table à l'autre à l'âge où

il migre d'une région à une autre. Le temps passé à risque par chaque individu est ainsi réparti entre les dénominateurs des taux des classes d'âge des différentes modalités du caractère qu'il a occupées pendant qu'il était à risque de vivre l'événement.

En pratique, ce travail se fait avec un logiciel d'analyse statistique. Règle générale, les données de l'enquête biographique sont mises à la disposition du chercheur sous la forme d'un fichier qui contient une ligne par individu ; c'est bien le cas de l'ESG de 2006 que nous utilisons ici. Chaque ligne contient toute l'information recueillie auprès d'un individu. Dans notre exemple, comme dans toute analyse, nous ne retenons de toute cette information que celle dont nous avons besoin : l'identifiant, le sexe, l'âge à l'enquête, l'âge au début de la vie au Canada, l'âge au moment de chaque changement de la situation conjugale — l'âge au début et à la fin de chaque union, l'âge à la transformation d'une union de fait en mariage — et, bien sûr, l'âge à la première naissance auquel on affecte une valeur élevée — ici 100 ans — lorsque la femme n'avait pas encore eu d'enfant au moment de l'enquête. La manière la plus générale de réaliser une analyse qui contient une VIFT consiste à préparer un fichier de données dans lequel la portion de la biographie de l'individu pendant laquelle il était à risque de changer d'état est découpée en plusieurs lignes, dont chacune correspond à la fraction du temps à risque située entre deux changements de modalité de la VIFT[9]. On prépare donc ici un fichier dans lequel chacune des lignes représente une fraction de la biographie d'une femme pendant laquelle celle-ci est à risque de donner naissance à son premier enfant, alors qu'elle vit au Canada, qui est située entre deux changements de sa situation conjugale.

Le tableau 4 présente un extrait de ce fichier de données, tel qu'il est préparé pour le logiciel Stata ; un autre logiciel représentera la même information autrement. La colonne intitulée « recid » contient l'identifiant, la colonne « SitConj », la situation conjugale — 0 lorsqu'on vit seul, 1 lorsqu'on vit avec son époux et 2 lorsqu'on vit en union de fait —, la colonne « AgeCan », l'âge au début de la vie au Canada et la colonne « AgeNaissıer », l'âge à la naissance du premier enfant. On trouve, dans la colonne « _to », le temps écoulé depuis l'origine de la table — l'âge de

9. Le *TDA* de Gotz Röhwer, basé sur le *Rate* de Nancy Tuma, de même que la procédure PHREG de *SAS* et l'instruction COXREG de *SPSS* opérationnalisent la VIFT en remplaçant la préparation des données sous forme de fichier biographique par un programme *ad hoc* qui est exécuté par le progiciel au moment de l'estimation.

TABLEAU 4 Extrait du fichier de données ayant servi à préparer le tableau 5 avant la division en classes d'âge.
Femmes âgées de 15 à 80 ans au moment de l'enquête

recid	_st	_to	_t	_d	SitConj	AgeCan	AgeNaiss1er
46	1	0	4,4	0	0	0	25,7
46	1	4,4	6,7	0	2	0	25,7
46	1	6,7	10,7	1	1	0	25,7
876	1	0	5,2	0	0	0	31,6
876	1	5,2	7,2	0	2	0	31,6
876	1	7,2	10,2	0	2	0	31,6
876	1	10,2	16,6	1	1	0	31,6
1173	1	7,5	8,4	0	0	22,5	100
1173	1	8,4	9,5	0	2	22,5	100
1173	1	9,5	15,4	0	1	22,5	100
1173	1	15,4	20,4	0	0	22,5	100
1173	1	20,4	23,9	0	2	22,5	100
2659	1	2,5	3	0	0	17,5	23,1
2659	1	3	3,6	0	2	17,5	23,1
2659	1	3,6	4,6	0	1	17,5	23,1
2659	1	4,6	7,2	0	0	17,5	23,1
2659	1	7,2	8,1	1	1	17,5	23,1

Source : Statistique Canada, Enquête sociale générale de 2006.

15 ans — au début de la fraction de la biographie représentée sur chaque ligne ; la colonne « _t » donne le temps écoulé depuis l'origine de la table à la fin de la fraction de la biographie représentée sur chaque ligne. La colonne « _st » indique si, oui ou non, la ligne correspond à une fraction de biographie pendant laquelle l'individu représenté sur cette ligne est à risque : dans certains cas, en effet, la préparation du fichier conduit à retenir la ligne qui correspond à un individu qui n'est jamais à risque ou à créer des lignes qui correspondent à des fractions de leur biographie où les individus ne sont pas à risque. La colonne « _d » indique la manière dont se termine la fraction de la biographie à laquelle correspond la ligne : la valeur 1 signifie que la fraction de biographie se termine par l'événement alors que la valeur 0 signifie le contraire. La valeur 0 est attribuée lorsque la fraction de biographie se termine par le passage d'une modalité à une autre d'une VIFT ou lorsqu'elle se termine parce que l'individu cesse d'être

observé, habituellement parce que la fin de cette fraction correspond au moment où il a été interrogé et qu'il n'a jamais vécu l'événement étudié avant ce moment. Ce détail est important : quitter le groupe à risque qui correspond à une modalité d'une VIFT pour passer à une autre modalité de la VIFT — par exemple cesser de vivre seul et commencer à vivre en union de fait — ne se distingue pas de quitter ce groupe à risque parce qu'on cesse d'être observé. Cette équivalence, qu'on voit ici dans le fichier de données préparé pour opérationnaliser la VIFT, vaut également dans les modèles statistiques eux-mêmes.

Les trois premières lignes du tableau 4 représentent la partie de la biographie d'une femme née au Canada pendant laquelle elle a été à risque de donner naissance à son premier enfant. Cette femme devient à risque de donner naissance à son premier enfant à 15 ans. La première ligne correspond à la partie de sa biographie qui précède la formation de sa première union, à l'âge de 19,4 ans. La seconde ligne correspond à l'union de fait qui commence à l'âge de 19,4 ans et se termine à 21,7 ans lorsqu'elle se marie. La troisième ligne correspond à la fraction de sa biographie qui commence au moment de son mariage et se termine par la naissance de son premier enfant, à l'âge de 25,7 ans, où elle cesse d'être à risque de donner naissance à cet enfant. Les quatre lignes suivantes proviennent d'une femme qui est née au Canada, qui a vécu en union de fait une première fois de 20,2 à 22,2 ans, puis une seconde fois de 22,2 à 25,2 ans ; elle s'est mariée à la fin de sa seconde union de fait, peut-être avec son conjoint du moment ; elle a eu son premier enfant à 31,6 ans alors qu'elle était toujours mariée. Les cinq lignes suivantes proviennent d'une femme qui s'est établie au Canada à 22,5 ans, qui a vécu en union de fait, s'est mariée puis s'est séparée ; au moment de l'enquête, elle vivait de nouveau en union de fait et n'avait pas encore donné naissance à son premier enfant. Les cinq dernières lignes proviennent d'une femme qui s'est établie au Canada à 17,5 ans, a vécu en union de fait de 18,0 à 18,6 ans, s'est mariée à 18,6 et a vécu avec son époux jusqu'à 19,6 ans, puis s'est mariée de nouveau (ou a repris la vie commune avec son premier époux) à 22,2 ans et a eu son premier enfant à 23,1 ans.

Chaque passage d'une modalité à une autre de la situation conjugale est un changement d'état qui définit une nouvelle ligne. Pour obtenir ces tableaux, il faut raisonner sur les classes d'âge comme on raisonne sur les modalités d'une VIFT : passer d'une classe d'âge à une autre, c'est passer d'une modalité à une autre d'une VIFT. En d'autres termes, il y a déjà une VIFT dans le tableau 2, même si, par économie d'espace, nous ne donnons pas d'échantillon du résultat de ce découpage dans le tableau 4. Pour le faire, il faudrait « découper », à chaque anniversaire, les lignes qui repré-

sentent la biographie de chaque femme de manière à marquer que le temps à risque entre le 15e et le 16e anniversaire est passé dans la première modalité de la VIFT qui représente l'âge, que le temps à risque entre le 16e et le 17e anniversaire est passé dans la deuxième modalité de la VIFT qui représente l'âge et ainsi de suite. Il est nécessaire de faire ce « découpage » pour estimer les modèles dont nous présentons les résultats, mais nous le laissons ici à l'imagination du lecteur[10].

Les ajouts et retraits du tableau 5 sont ceux du tableau 2 — respectivement l'établissement des immigrantes au Canada et la sortie de l'observation au moment de l'enquête des femmes qui n'avaient pas encore eu leur premier enfant au moment de l'enquête — auxquels s'ajoutent ceux qui résultent des déplacements entre les modalités de la situation conjugale. Le temps à risque est réparti de manière fine entre les modalités lorsque ce déplacement se fait en cours d'intervalle, ce qui est le cas le plus fréquent. Les ajouts et retraits qui résultent de déplacements entre les modalités de la VIFT sont tout à fait analogues à ceux qui résultent des déplacements entre les régions d'une table multirégionale. Dit autrement, l'ajout des immigrants selon la logique des entrées échelonnées, la sortie de l'observation des femmes qui n'avaient pas encore donné naissance à leur premier enfant au moment de l'enquête et les déplacements entre les modalités d'une VIFT se ramènent tous à ajouter ou retirer des individus du groupe à risque de la classe d'âge appropriée.

Dans notre exemple, la femme ne peut donner naissance qu'une seule fois : cet événement sera compté dans le numérateur du taux de la classe d'âge qui correspond à l'âge où elle a eu son premier enfant dans la table qui correspond à la modalité de la situation conjugale qui était la sienne au moment où elle a donné naissance.

Le risque est celui auquel sont soumises toutes les femmes qui se trouvent dans une situation conjugale donnée et dans une classe d'âge donnée, peu importe l'âge auquel elles sont entrées dans cette situation conjugale. Ainsi, toutes les femmes mariées âgées de 20 à 21 ans qui n'ont pas déjà donné naissance à leur premier enfant sont soumises au même risque, peu

10. Ce fait peut paraître curieux, mais il n'est pas rare. Tous les modèles paramétriques par parties découpent le temps de cette manière pour estimer un quotient instantané dont la valeur est constante à l'intérieur de chaque partie. En microsimulation, il est d'usage d'utiliser des quotients instantanés constants à l'intérieur de parties comme les classes d'âge ou encore des classes de temps écoulé depuis un changement d'état, p. ex. le changement de situation conjugale le plus récent ou la naissance du dernier enfant.

TABLEAU 5A La table d'extinction comme modèle statistique, selon une variable indépendante fonction du temps – L'âge à la naissance du premier enfant selon l'union et son type. Hors union – Femmes âgées de 15 à 80 ans au moment de l'enquête

T	À RISQUE AU DÉBUT	AJOUTS	RETRAITS	ÉVÉNEMENTS	r_t	m_t	q_t	*f(t)*	*F(t)*	*S(t)*	*H(t)*
De 15 ans à moins de 16	10 980	6	229	25	10 871,30	0,0023	0,0023	0,0023	0,0023	0,9977	0,0023
De 16 ans à moins de 17	10 732	13	401	79	10 539,20	0,0075	0,0075	0,0075	0,0098	0,9902	0,0098
De 17 ans à moins de 18	10 265	231	599	119	10 073,85	0,0118	0,0117	0,0116	0,0214	0,9786	0,0216
De 18 ans à moins de 19	9 778	37	919	124	9 330,70	0,0133	0,0132	0,0129	0,0343	0,9657	0,0349
De 19 ans à moins de 20	8 772	62	1 105	137	8 243,90	0,0166	0,0165	0,0159	0,0502	0,9498	0,0515
De 20 ans à moins de 21	7 592	82	1 126	128	7 060,40	0,0181	0,0179	0,017	0,0673	0,9327	0,0696
De 21 ans à moins de 22	6 420	87	1 103	90	5 920,75	0,0152	0,0151	0,0141	0,0813	0,9187	0,0848
De 22 ans à moins de 23	5 314	299	955	89	4 994,95	0,0178	0,0176	0,0162	0,0975	0,9025	0,1026
De 23 ans à moins de 24	4 569	114	873	55	4 186,35	0,0131	0,013	0,0117	0,1093	0,8907	0,1157
De 24 ans à moins de 25	3 755	112	680	48	3 454,15	0,0139	0,0138	0,0123	0,1216	0,8784	0,1296
De 25 ans à moins de 26	3 139	99	589	44	2 905,75	0,0151	0,015	0,0132	0,1347	0,8653	0,1447
De 26 ans à moins de 27	2 605	89	420	15	2 442,10	0,0061	0,0061	0,0053	0,14	0,86	0,1508
De 27 ans à moins de 28	2 259	199	385	42	2 153,10	0,0195	0,0193	0,0166	0,1566	0,8434	0,1703
De 28 ans à moins de 29	2 031	68	340	18	1 891,65	0,0095	0,0094	0,008	0,1646	0,8354	0,1798
De 29 ans à moins de 30	1 741	87	255	18	1 654,90	0,0109	0,0108	0,0091	0,1736	0,8264	0,1907
De 30 ans à moins de 31	1 555	72	226	14	1 473,10	0,0095	0,0094	0,0078	0,1814	0,8186	0,2002
De 31 ans à moins de 32	1 387	48	178	17	1 313,55	0,0129	0,0128	0,0105	0,1919	0,8081	0,2131
De 32 ans à moins de 33	1 240	98	143	18	1 211,75	0,0149	0,0148	0,012	0,2039	0,7961	0,228
De 33 ans à moins de 34	1 177	46	131	7	1 129,25	0,0062	0,0062	0,0049	0,2088	0,7912	0,2342
De 34 ans à moins de 35	1 085	41	104	7	1047,60	0,0067	0,0067	0,0053	0,2141	0,7859	0,2409
De 35 ans à moins de 36	1 015	43	114	4	980,35	0,0041	0,0041	0,0032	0,2173	0,7827	0,245
De 36 ans à moins de 37	940	30	106	6	899,95	0,0067	0,0067	0,0052	0,2225	0,7775	0,2517
De 37 ans à moins de 38	858	59	83	5	843,65	0,0059	0,0059	0,0046	0,2271	0,7729	0,2576
De 38 ans à moins de 39	829	29	74	4	808,75	0,0049	0,0049	0,0038	0,2309	0,7691	0,2625
De 39 ans à moins de 40	780	28	805	3	756,60	0,004	0,004	0,0031	0,2339	0,7661	0,2665
Somme				1 116							

Source : Statistique Canada, Enquête sociale générale de 2006.

TABLEAU 5B La table d'extinction comme modèle statistique, selon une variable indépendante fonction du temps – L'âge à la naissance du premier enfant selon l'union et son type. Mariage – Femmes âgées de 15 à 80 ans au moment de l'enquête

T	À RISQUE AU DÉBUT	AJOUTS	RETRAITS	ÉVÉNEMENTS	r_t	m_t	q_t	*f(t)*	*F(t)*	*S(t)*	*H(t)*
De 15 ans à moins de 16	33	51	0	16	45,70	0,3501	0,2904	0,2904	0,2904	0,7096	0,2904
De 16 ans à moins de 17	68	132	0	45	94,30	0,4772	0,3738	0,2652	0,5556	0,4444	0,7676
De 17 ans à moins de 18	155	324	4	140	227,10	0,6165	0,4538	0,2017	0,7573	0,2427	1,3841
De 18 ans à moins de 19	335	557	8	240	466,20	0,5148	0,4013	0,0974	0,8547	0,1453	1,8989
De 19 ans à moins de 20	644	734	15	354	790,50	0,4478	0,3604	0,0524	0,9071	0,0929	2,3467
De 20 ans à moins de 21	1 009	741	22	463	1 130,55	0,4095	0,3403	0,0316	0,9387	0,0613	2,7562
De 21 ans à moins de 22	1 265	764	32	412	1 411,30	0,2919	0,2557	0,0157	0,9544	0,0456	3,0481
De 22 ans à moins de 23	1 585	737	39	459	1 690,15	0,2716	0,2406	0,011	0,9653	0,0347	3,3197
De 23 ans à moins de 24	1 824	631	53	461	1 892,10	0,2436	0,2184	0,0076	0,9729	0,0271	3,5633
De 24 ans à moins de 25	1 941	525	63	450	1 958,30	0,2298	0,2077	0,0056	0,9785	0,0215	3,7931
De 25 ans à moins de 26	1 953	387	69	489	1 868,85	0,2617	0,2321	0,005	0,9835	0,0165	4,0548
De 26 ans à moins de 27	1 782	318	68	395	1 711,10	0,2308	0,2077	0,0034	0,9869	0,0131	4,2856
De 27 ans à moins de 28	1 637	353	66	384	1 609,95	0,2385	0,216	0,0028	0,9898	0,0102	4,5241
De 28 ans à moins de 29	1 540	240	58	309	1 489,85	0,2074	0,19	0,0019	0,9917	0,0083	4,7315
De 29 ans à moins de 30	1 413	193	64	309	1 336,30	0,2312	0,2091	0,0017	0,9934	0,0066	4,9627
De 30 ans à moins de 31	1 233	147	66	220	1 151,85	0,191	0,1733	0,0011	0,9946	0,0054	5,1537
De 31 ans à moins de 32	1 094	129	38	180	1 057,05	0,1703	0,1577	0,0009	0,9954	0,0046	5,324
De 32 ans à moins de 33	1 005	108	49	151	956,85	0,1578	0,1459	0,0007	0,9961	0,0039	5,4818
De 33 ans à moins de 34	913	71	40	117	868,95	0,1346	0,1257	0,0005	0,9966	0,0034	5,6164
De 34 ans à moins de 35	827	64	35	84	801,80	0,1048	0,0994	0,0003	0,9969	0,0031	5,7212
De 35 ans à moins de 36	772	64	42	67	751,45	0,0892	0,0856	0,0003	0,9972	0,0028	5,8104
De 36 ans à moins de 37	727	43	30	53	713,45	0,0743	0,0719	0,0002	0,9974	0,0026	5,8847
De 37 ans à moins de 38	687	49	28	43	678,35	0,0634	0,0615	0,0002	0,9976	0,0024	5,9481
De 38 ans à moins de 39	665	24	31	19	649,15	0,0293	0,0288	0,0001	0,9976	0,0024	5,9774
De 39 ans à moins de 40	639	23	647	15	621,54	0,0241	0,0238	0,0001	0,9977	0,0023	6,0015
Somme				5 875							

Source : Statistique Canada, Enquête sociale générale de 2006.

TABLEAU 5C La table d'extinction comme modèle statistique, selon une variable indépendante fonction du temps – L'âge à la naissance du premier enfant selon l'union et son type. Union de fait – Femmes âgées de 15 à 80 ans au moment de l'enquête

T	À RISQUE AU DÉBUT	AJOUTS	RETRAITS	ÉVÉNEMENTS	r_t	m_t	q_t	*f(t)*	*F(t)*	*S(t)*	*H(t)*
De 15 ans à moins de 16	25	36	8	2	33,70	0,0593	0,0588	0,0588	0,0588	0,9412	0,0588
De 16 ans à moins de 17	51	122	16	8	91,40	0,0875	0,0841	0,0792	0,1379	0,8621	0,1463
De 17 ans à moins de 18	149	169	36	27	190,65	0,1416	0,1328	0,1145	0,2524	0,7476	0,2879
De 18 ans à moins de 19	255	273	62	42	343,7	0,1222	0,1161	0,0868	0,3392	0,6608	0,4101
De 19 ans à moins de 20	424	323	110	67	489,45	0,1369	0,1281	0,0847	0,4239	0,5761	0,547
De 20 ans à moins de 21	570	336	138	62	634,6	0,0977	0,0934	0,0538	0,4777	0,5223	0,6447
De 21 ans à moins de 22	706	323	168	66	743,9	0,0887	0,0852	0,0445	0,5222	0,4778	0,7334
De 22 ans à moins de 23	795	325	197	66	833,35	0,0792	0,0761	0,0363	0,5585	0,4415	0,8126
De 23 ans à moins de 24	857	292	218	81	862,3	0,0939	0,0902	0,0398	0,5983	0,4017	0,9065
De 24 ans à moins de 25	850	229	233	43	838,35	0,0513	0,0501	0,0201	0,6184	0,3816	0,9578
De 25 ans à moins de 26	803	235	204	70	782,95	0,0894	0,086	0,0328	0,6512	0,3488	1,0472
De 26 ans à moins de 27	764	141	166	37	735,9	0,0503	0,0491	0,0171	0,6684	0,3316	1,0975
De 27 ans à moins de 28	702	151	171	47	676	0,0695	0,0672	0,0223	0,6907	0,3093	1,167
De 28 ans à moins de 29	635	126	155	45	601,35	0,0748	0,0719	0,0222	0,7129	0,2871	1,2418
De 29 ans à moins de 30	561	97	144	27	525,2	0,0514	0,0501	0,0144	0,7273	0,2727	1,2932
De 30 ans à moins de 31	487	104	113	44	463,55	0,0949	0,0908	0,0248	0,752	0,248	1,3881
De 31 ans à moins de 32	434	62	93	31	407,1	0,0761	0,0737	0,0183	0,7703	0,2297	1,4642
De 32 ans à moins de 33	372	63	76	24	359,95	0,0667	0,0647	0,0149	0,7852	0,2148	1,5309
De 33 ans à moins de 34	335	65	61	25	328,8	0,076	0,0736	0,0158	0,801	0,199	1,6069
De 34 ans à moins de 35	314	38	58	18	296,9	0,0606	0,0589	0,0117	0,8127	0,1873	1,6675
De 35 ans à moins de 36	276	47	58	8	268,7	0,0298	0,0294	0,0055	0,8182	0,1818	1,6973
De 36 ans à moins de 37	257	33	32	11	250,1	0,044	0,043	0,0078	0,826	0,174	1,7413
De 37 ans à moins de 38	247	23	41	4	234,2	0,0171	0,017	0,003	0,829	0,171	1,7584
De 38 ans à moins de 39	225	33	38	7	218,75	0,032	0,0315	0,0054	0,8343	0,1657	1,7904
De 39 ans à moins de 40	213	12	220	5	199,57	0,0251	0,0247	0,0041	0,8384	0,1616	1,8155
Somme				867							

Source : Statistique Canada, Enquête sociale générale de 2006.

importe qu'elles viennent tout juste de se marier ou qu'elles se soient mariées à 15 ans. La fonction de survie donne la proportion des femmes mariées à 15 ans qui n'auraient toujours pas eu leur premier enfant à la fin de chaque intervalle, même si le risque, au cours de chacun des intervalles qui suivent le premier, est principalement calculé à partir du temps à risque et des naissances de femmes qui se sont mariées après l'âge de 15 ans.

La VIFT au sens strict dans les modèles linéaires

À strictement parler, ce que nous venons de présenter n'est pas une VIFT, mais plutôt ce qui, dans les modèles de risque, se nomme la stratification. Nous avons estimé trois séries de valeurs du quotient instantané qui ne sont pas reliées entre elles, une pour chacune des modalités de la situation conjugale. La VIFT est très semblable à la stratification, mais à une différence près : plutôt que d'estimer trois séries de valeurs qui ne sont pas reliées entre elles, on estime d'une part une seule série de valeurs « moyennes » et d'autre part le rapport « moyen » à cette série pour chacune des modalités de la VIFT. En d'autres mots, dans notre exemple, la VIFT au sens strict remplace les trois séries de valeurs par une seule série et deux coefficients. La souplesse de la stratification est remplacée par la rigidité de la proportionnalité, telle qu'elle apparaît dans l'équation par laquelle on représente le plus souvent les modèles de risques proportionnels, c'est-à-dire $h(t) = h_0(t)\exp(x\beta)$, sous sa forme multiplicative et $\ln[h(t)] = \ln[h_0(t)] + x\beta$, sous sa forme additive. L'écart entre la série des logarithmes des taux associés au fait de ne pas vivre en union et la série des logarithmes des taux associés au mariage est toujours le même, comme l'écart entre la série des logarithmes des taux associés au fait de ne pas vivre en union et la série des logarithmes des taux associés à l'union libre est toujours le même.

Nous avons écrit les mots « moyennes » et « moyen » entre guillemets parce que ce que nous venons d'écrire est parfaitement exact lorsqu'on représente les modalités de la VIFT au moyen du codage utilisé de manière habituelle en analyse de la variance — où la modalité de référence est représentée par une suite de –1 —, mais n'est pas tout à fait exact lorsqu'on représente ces modalités comme on le fait généralement en régression, où la modalité de référence est représentée par une suite de 0. Lorsqu'on utilise le codage habituel de la régression, la série de valeurs associée à la modalité de référence est plutôt égale au produit de la série moyenne et du rapport moyen à la série moyenne de la modalité de référence, alors que le

coefficient associé à chacune des autres modalités est le produit du rapport moyen de cette modalité et de l'inverse du rapport moyen de la modalité de référence. Ceci est vrai pour tous les modèles de risques proportionnels.

On ne peut pas estimer, avec une table d'extinction, l'effet d'une « vraie » VIFT au sens où nous la décrivons au paragraphe précédent. On ne peut estimer l'effet d'une VIFT qu'au moyen d'un modèle statistique et, pour des raisons qui deviendront progressivement évidentes, on ne peut le faire que lorsque ce modèle est estimé par la méthode du maximum de vraisemblance ou une méthode qui lui est apparentée.

La méthode du maximum de vraisemblance est la solution que Fisher (1922) a proposée au problème de la probabilité inverse. On peut la présenter directement à partir d'un cas semblable à celui qui nous occupe. On formule une équation qui relie une variable dépendante à une ou plusieurs variables indépendantes. On ne connaît pas la valeur des coefficients qui sont associés aux variables indépendantes et on veut trouver les valeurs de ces coefficients qui fassent en sorte que la probabilité d'avoir échantillonné les valeurs qu'on a échantillonnées soit la plus élevée. Il a été démontré que ce problème, posé dans ces termes, est insoluble. Il est cependant possible de le résoudre en remplaçant la probabilité par une quantité qui lui est reliée de manière directe et qu'on nomme la *vraisemblance* : on maximise alors la probabilité en maximisant la vraisemblance. La *vraisemblance d'un échantillon* — généralement notée L ou $\mathcal{L}$, de l'anglais « *likelihood* » — est le produit de la vraisemblance de chacune des observations qui le composent. La *vraisemblance d'une observation*[11] est elle-même une équation construite à partir de deux éléments : l'expression algébrique d'une des fonctions de la loi de probabilité que suit la variable aléatoire — habituellement sa fonction de densité de probabilité — et l'expression algébrique de la relation entre la variable dépendante et les variables indépendantes — habituellement la relation entre la valeur prédite de la variable dépendante et les variables indépendantes, qui est habituellement la somme des effets linéaires de chacune de celles-ci. Dans le cas le plus simple, un des paramètres de la loi de probabilité correspond à l'aspect de la variable dépendante qui est relié aux variables indépendantes ; on obtient alors la vraisemblance de l'observation en remplaçant ce paramètre par la combinaison des variables indépendantes qui corres-

11. Rappelons qu'en statistique, le mot « observation » ne désigne jamais l'*unité statistique* (p. ex. l'individu échantillonné dans une population *finie* ou « tiré » d'une population *finie*), mais bien le résultat d'un tirage dans une population *théorique*.

pond à cet aspect de la variable dépendante. Dans le cas de régression multiple, on pose que le modèle suit une loi normale et on obtient la vraisemblance de l'unité en remplaçant la moyenne qui apparaît dans l'expression algébrique de la fonction de densité de la loi normale, par la somme des effets des variables indépendantes qui correspond à la valeur prédite de la régression multiple. L'estimation proprement dite se fait par approximations successives, en cherchant les valeurs des coefficients associés à chacune des variables indépendantes qui maximisent le produit de toutes les vraisemblances des observations et donc la vraisemblance de l'échantillon. À la première étape, on calcule la valeur de la vraisemblance de chaque observation en imposant une valeur arbitraire à chacun des coefficients à estimer, très souvent zéro. On calcule le produit des vraisemblances (en pratique, la somme de leurs logarithmes), on calcule les dérivées partielles première et seconde (du logarithme) de la fonction de vraisemblance par rapport à chacun des coefficients et, en supposant que la fonction de vraisemblance soit convexe, on se base sur les valeurs de ces dérivées pour choisir la nouvelle valeur de chacun des coefficients. On répète ces opérations jusqu'à ce que la valeur de la vraisemblance (et de son logarithme) ait atteint un maximum qu'il ne semble plus possible de dépasser. Chacune de ces étapes se nomme « itération », la quantité qu'on maximise est le logarithme de la vraisemblance (généralement noté l ou ℓ) et la technique de recherche par approximations successives que nous venons de décrire est connue sous le nom d'« algorithme de Newton-Raphson ».

La vraisemblance d'une observation est une fonction de la probabilité de l'observer. Règle générale, on en fait donc une fonction de la fonction de densité de probabilité de la loi de probabilité du modèle (*sic*). Cette règle ne vaut pas lorsque la distribution des fréquences ne correspond pas à la fonction de densité de la variable aléatoire dont elle est tirée, ce qui est le cas ici comme nous l'avons vu en examinant les distributions du tableau 1. On résout le problème en suivant un raisonnement analogue à celui qui fonde la table d'extinction. La vraisemblance de chaque observation devient ainsi une fonction de la fonction de risque. En plus de résoudre le problème, cette solution a un avantage considérable : elle permet d'intégrer les VIFT au modèle.

Le moment où une unité statistique change d'état ou cesse d'être observée est une constante pour cette unité : chaque femme qui a eu son premier enfant l'a eu à l'âge où elle l'a eu. Par contre, pour chaque femme, la probabilité d'avoir le premier enfant peut, en principe, varier d'un

instant à l'autre tant qu'elle ne l'a pas eu ; il en est de même de la probabilité de ne pas avoir encore eu le premier enfant, qui varie elle aussi d'un instant à l'autre. Autrement dit, la durée dans l'état d'origine est une constante pour chaque unité statistique, alors que la probabilité de changer d'état peut varier d'un instant à l'autre et que la probabilité de ne pas avoir encore changé d'état varie elle aussi d'un instant à l'autre. La *durée* (la valeur t de la variable T pour chaque unité statistique) correspond à la fonction de densité de la variable T, la *probabilité instantanée* de changer d'état (le quotient instantané associé à chaque valeur t de T) correspond à la fonction de risque de la variable T et la *probabilité de ne pas avoir encore changé d'état* au temps t correspond à la fonction de survie de la variable T. Construire la vraisemblance de l'échantillon à partir de la fonction de densité impose d'avoir exactement un terme par unité statistique et ne permet donc pas de tenir compte de plus d'une valeur par variable indépendante par unité statistique. Construire la vraisemblance à partir de la fonction de risque permet de ne pas lier le nombre des termes au nombre des unités statistiques et permet de tenir compte des valeurs des variables indépendantes de plusieurs unités statistiques dans chaque terme. Ces deux propriétés permettent d'intégrer, à la fonction de vraisemblance, plus d'une valeur par variable indépendante par unité statistique, ce qui rend possible l'usage de VIFT. La vraisemblance des modèles de risque est ainsi construite à partir de la fonction de risque entendue comme le rapport de la fonction de densité à la fonction de survie.

Le modèle de Cox se distingue des modèles paramétriques en ce qu'il n'utilise pas de loi de probabilité ou, plus précisément, en ce qu'il utilise une loi de probabilité non paramétrique. Dans le modèle de Cox, la loi de probabilité paramétrique — c'est-à-dire qu'on peut exprimer de manière algébrique — est remplacée par la loi de probabilité qu'on obtient en utilisant l'estimateur de Kaplan-Meier, dont ces deux auteurs ont montré qu'il produisait la meilleure estimation possible, au sens du maximum de vraisemblance, de la loi de probabilité qui régit un changement d'état (Kaplan et Meier, 1958). La vraisemblance basée sur cet estimateur de la loi de probabilité est dite « partielle » parce qu'elle sert à maximiser une quantité qui dépend des coefficients associés aux variables indépendantes (notés $\boldsymbol{\beta}$), mais pas des paramètres de la loi de probabilité qui en est justement dépourvue.

On peut représenter comme suit l'équation de vraisemblance partielle du modèle de Cox, où *PL* désigne la vraisemblance partielle :

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^{n} \left(\frac{\exp(\mathrm{x}_t \boldsymbol{\beta})}{\sum_{i=1}^{n} \exp(\mathrm{x}_{it} \boldsymbol{\beta}) \cdot r_{it}} \right)^{d_t} .$$

Cette équation se comprend mieux si on imagine que les unités statistiques, au nombre de *n*, sont rangées, en ordre croissant de *t*, dans l'ordre où elles cessent d'être observées, peu importe que ce soit parce qu'elles changent d'état ou non. La variable d_t vaut 1 si l'unité cesse d'être observée parce qu'elle change d'état et 0 autrement. On présume qu'il est possible de ranger les unités de manière parfaite, aucune ne cessant d'être observée au même moment qu'une autre. On a, en principe, un terme pour chaque unité, mais ce terme n'affecte pas la valeur de la vraisemblance de l'échantillon si elle correspond à une unité qui cesse d'être observée sans changer d'état : dans ce cas, d_t vaut 0, le terme est élevé à la puissance 0 et vaut ainsi 1, l'élément neutre de la multiplication. Le rapport entre quantités qui se trouve au cœur de chaque contribution est construit comme le rapport qui se trouve au cœur de l'estimateur de Kaplan-Meier et qui y joue le rôle du quotient d'une table d'extinction conventionnelle. Dans l'estimateur de Kaplan-Meier, ce terme est formé, au numérateur, du nombre 1 — qui est le seul nombre de changements d'état que cet estimateur admet en principe au cours d'un intervalle — et, au dénominateur, du nombre des unités encore à risque de changer d'état au moment où change d'état l'unité qui termine l'intervalle et justifie le calcul du terme. Dans le modèle de Cox, le terme équivalent est formé, au numérateur, de la combinaison des effets des variables indépendantes de l'unité statistique qui change d'état au temps *t* et, au dénominateur, de la somme des combinaisons des valeurs des variables indépendantes des unités qui sont encore à risque de changer d'état au temps *t*. La variable r_{it} vaut 1 si l'unité *i* est à risque au temps *t* et 0 autrement, et indique qu'on ne tient pas compte des valeurs des variables indépendantes des unités qui ne sont plus observées au temps *t*. Chaque terme de la fonction de vraisemblance du modèle de Cox intègre donc la valeur, au moment où elle change d'état, de chaque VIFT de l'unité statistique qui change d'état, ainsi que la valeur, à ce moment, de chaque VIFT de toutes les unités statistiques à risque au moment de ce changement d'état. L'effet d'une VIFT représente l'écart moyen, au quotient

instantané moyen, de chacune des modalités de la VIFT, exprimé sous forme de rapport à l'écart moyen au quotient instantané moyen de la modalité de référence, même si le quotient instantané ne peut pas être estimé dans le cas du modèle de Cox, bien qu'il existe en principe.

Dans le calcul de la vraisemblance de l'échantillon, un épisode peut, en principe, être découpé en un nombre infini de segments dont seul le dernier peut se terminer par un événement. Cette propriété permet les entrées échelonnées, l'hiatus — c'est-à-dire, au sein d'un seul épisode, la sortie suivie du retour dans le groupe à risque — ainsi que les variables indépendantes fonction du temps. Ce découpage se manifeste dans le jeu des ajouts et des retraits au groupe à risque.

On peut représenter comme suit l'équation de vraisemblance des modèles de risque paramétriques :

$$L(\boldsymbol{\beta},\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{j=1}^{k_i} \frac{S\left(t_{ij}\middle|\mathbf{x}_{ij}\boldsymbol{\beta},\boldsymbol{\theta}\right)^{1-d_{ij}} f\left(t_{ij}\middle|\mathbf{x}_{ij}\boldsymbol{\beta},\boldsymbol{\theta}\right)^{d_{ij}}}{S\left(t_{0ij}\middle|\mathbf{x}_{ij}\boldsymbol{\beta},\boldsymbol{\theta}\right)}.$$

Cette représentation est très générale. $L(\boldsymbol{\beta},\boldsymbol{\theta})$ indique que la vraisemblance L est fonction à la fois des coefficients associés aux variables indépendantes, qui forment le vecteur $\boldsymbol{\beta}$, et des paramètres de la loi de probabilité qui régit la variable aléatoire, qui forment le vecteur $\boldsymbol{\theta}$. L'échantillon est formé de n unités statistiques. L'épisode de chaque unité i est découpé en k_i segments. Dans le cas le plus simple, aucune des variables indépendantes de cette unité ne change de valeur en fonction du temps et l'épisode n'a qu'un segment ; ce segment commence à t_{0i1} et se termine à t_{in}. Dans le cas le plus habituel, la valeur d'une ou plusieurs des variables indépendantes du vecteur $\mathbf{x}$ de cette unité change en fonction du temps et l'épisode de cette unité est découpé en deux ou plusieurs segments. Le premier segment commence à t_{0i1} et se termine à t_{in}, le deuxième segment commence à t_{0i2} et se termine à t_{i2} et ainsi de suite jusqu'au dernier segment qui commence à t_{0ik_i} et se termine à t_{ik_i}. Le premier segment commence au début de l'épisode et le dernier se termine à la fin de l'épisode. Chaque segment correspond à une fraction de l'épisode pendant laquelle les valeurs des variables indépendantes du vecteur $\mathbf{x}_{ij}$ ne changent pas. Règle générale, le nombre des segments k_i est égal au nombre des changements de valeur des variables indépendantes au cours de l'épisode plus un ; le nombre des segments peut être plus petit si la valeur de deux ou plusieurs variables indépendantes change au même moment. Tout ceci

revient à reprendre de manière formelle la description des lignes du tableau 4.

De manière analogue à ce que nous avons vu dans la fonction de vraisemblance du modèle de Cox, la variable d_{ij} vaut 1 si l'unité cesse d'être observée à la fin d'un segment parce qu'elle change d'état et 0 autrement. Le terme qui correspond à un segment vaut donc le rapport entre la valeur de la fonction de densité à la fin du segment et la valeur de la fonction de survie au début du segment lorsque le segment se termine par l'événement, et vaut le rapport entre la valeur de la fonction de survie à la fin du segment et la valeur de la fonction de survie au début du segment dans tous les autres cas. En d'autres mots, le terme vaut le risque au moment de l'événement lorsque le segment se termine par un événement puisque $h(t) = f(t) / S(t)$. Dans tous les cas, la valeur de la densité et la valeur de la survie sont évaluées en fonction des valeurs des variables indépendantes au cours du segment. C'est de cette manière qu'on intègre les VIFT dans les modèles de risque paramétriques.

On obtient la fonction de vraisemblance d'un modèle paramétrique précis en remplaçant l'expression générale de la fonction de densité et de la fonction de survie par les formes algébriques appropriées. Sachant que la fonction de densité et la fonction de survie de la loi exponentielle sont respectivement $f(t) = \lambda \exp(-\lambda t)$ et $S(t) = \exp(-\exp(\lambda t))$ et sachant par ailleurs que dans le modèle exponentiel, on pose que le paramètre *?* est la valeur prédite, c'est-à-dire $\hat{\lambda}_{ij} = \lambda_0 + \mathbf{x}_{ij}\boldsymbol{\beta}$, on obtient :

$$L(\boldsymbol{\beta}, \lambda_0) = \prod_{i=1}^{n} \prod_{j=1}^{k_i} \frac{\left(\exp\left(-\exp\left(\lambda_0 + \mathbf{x}_{ij}\boldsymbol{\beta}\right)t_{ij}\right)\right)^{1-d_{ij}} \left(\left(\lambda_0 + \mathbf{x}_{ij}\boldsymbol{\beta}\right)\exp\left(-\left(\lambda_0 + \mathbf{x}_{ij}\boldsymbol{\beta}\right)t_{ij}\right)\right)^{d_{ij}}}{\left(\exp\left(-\exp\left(\lambda_0 + \mathbf{x}_{ij}\boldsymbol{\beta}\right)t_{0ij}\right)\right)}$$

qui achève d'illustrer comment on intègre les valeurs des VIFT dans l'équation de vraisemblance d'un modèle de risque paramétrique.

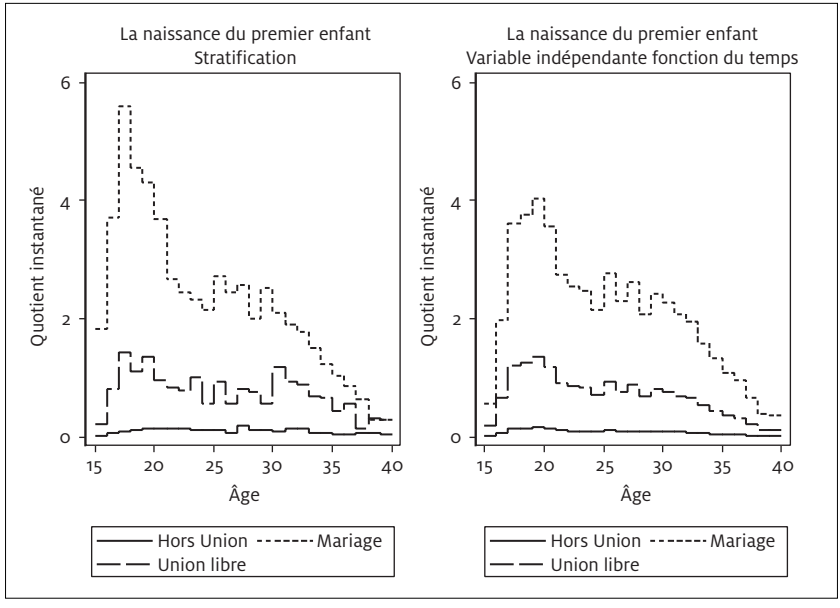
Les trois premières colonnes du tableau 6 correspondent aux colonnes m_t des tableaux 5a, 5b et 5c, les taux étant estimés cette fois-ci en tenant compte des poids de sondage. Les trois colonnes suivantes donnent les quotients instantanés obtenus au moyen du modèle de Poisson par parties stratifié ; les colonnes qui suivent donnent les quotients instantanés obtenus au moyen du modèle exponentiel par parties stratifié. L'avant-dernière colonne donne les résultats du modèle de Poisson par parties dans lequel la VIFT est construite de manière stricte ; la dernière colonne donne les résultats du modèle exponentiel par parties dans lequel la VIFT est là aussi construite de manière stricte.

TABLEAU 6 La stratification et la VIFT. Comparaison de la table et de deux modèles paramétriques par parties – L'âge à la naissance du premier enfant selon l'union et son type. Estimation pondérée – Femmes âgées de 15 à 80 ans au moment de l'enquête

	Table (m_t)			Poisson			Exponentiel			Poisson	Exponentiel
T	Hors union	Mariage	Union libre	Hors union	Mariage	Union libre	Hors union	Mariage	Union libre		
De 15 ans à moins de 16	0,0017	0,1838	0,0208	0,0017	0,1838	0,0208	0,0017	0,1838	0,0208	0,0022	0,0022
De 16 ans à moins de 17	0,0062	0,3704	0,0816	0,0062	0,3704	0,0816	0,0062	0,3704	0,0816	0,0077	0,0077
De 17 ans à moins de 18	0,0097	0,5606	0,1425	0,0097	0,5606	0,1425	0,0097	0,5606	0,1425	0,0141	0,0141
De 18 ans à moins de 19	0,0118	0,4558	0,1107	0,0118	0,4558	0,1107	0,0118	0,4558	0,1107	0,0147	0,0147
De 19 ans à moins de 20	0,0134	0,4316	0,1355	0,0134	0,4316	0,1355	0,0134	0,4316	0,1355	0,0158	0,0158
De 20 ans à moins de 21	0,0145	0,3676	0,096	0,0145	0,3676	0,096	0,0145	0,3676	0,096	0,0139	0,0139
De 21 ans à moins de 22	0,013	0,2664	0,0843	0,013	0,2664	0,0843	0,013	0,2664	0,0843	0,0107	0,0107
De 22 ans à moins de 23	0,0152	0,2458	0,0779	0,0152	0,2458	0,0779	0,0152	0,2458	0,0779	0,01	0,01
De 23 ans à moins de 24	0,0125	0,2329	0,1	0,0125	0,2329	0,1	0,0125	0,2329	0,1	0,0097	0,0097
De 24 ans à moins de 25	0,0119	0,2147	0,0562	0,0119	0,2147	0,0562	0,0119	0,2147	0,0562	0,0084	0,0084
De 25 ans à moins de 26	0,0122	0,2731	0,0942	0,0122	0,2731	0,0942	0,0122	0,2731	0,0942	0,0108	0,0108
De 26 ans à moins de 27	0,0065	0,244	0,0556	0,0065	0,244	0,0556	0,0065	0,244	0,0556	0,009	0,009
De 27 ans à moins de 28	0,0181	0,2578	0,0805	0,0181	0,2578	0,0805	0,0181	0,2578	0,0805	0,0103	0,0103
De 28 ans à moins de 29	0,0108	0,2005	0,0754	0,0108	0,2005	0,0754	0,0108	0,2005	0,0754	0,0081	0,0081
De 29 ans à moins de 30	0,0118	0,2514	0,0553	0,0118	0,2514	0,0553	0,0118	0,2514	0,0553	0,0095	0,0095
De 30 ans à moins de 31	0,0096	0,2098	0,1181	0,0096	0,2098	0,1181	0,0096	0,2098	0,1181	0,0089	0,0089
De 31 ans à moins de 32	0,0144	0,1909	0,0935	0,0144	0,1909	0,0935	0,0144	0,1909	0,0935	0,0081	0,0081
De 32 ans à moins de 33	0,0148	0,1778	0,0897	0,0148	0,1778	0,0897	0,0148	0,1778	0,0897	0,0076	0,0076
De 33 ans à moins de 34	0,0074	0,1509	0,0681	0,0074	0,1509	0,0681	0,0074	0,1509	0,0681	0,0062	0,0062
De 34 ans à moins de 35	0,0078	0,1227	0,0673	0,0078	0,1227	0,0673	0,0078	0,1227	0,0673	0,0052	0,0052
De 35 ans à moins de 36	0,0051	0,1046	0,0433	0,0051	0,1046	0,0433	0,0051	0,1046	0,0433	0,0042	0,0042
De 36 ans à moins de 37	0,0039	0,0873	0,057	0,0039	0,0873	0,057	0,0039	0,0873	0,057	0,0038	0,0038
De 37 ans à moins de 38	0,0064	0,0646	0,0153	0,0064	0,0646	0,0153	0,0064	0,0646	0,0153	0,0026	0,0026
De 38 ans à moins de 39	0,0063	0,0289	0,0305	0,0063	0,0289	0,0305	0,0063	0,0289	0,0305	0,0015	0,0015
De 39 ans à moins de 40	0,0036	0,0294	0,0292	0,0036	0,0294	0,0292	0,0036	0,0294	0,0292	0,0014	0,0014
Hors union										1	1
Mariage										25,5536	25,5536
Union de fait										8,6003	8,6003

Source : Statistique Canada, Enquête sociale générale de 2006.

FIGURE 1 La stratification et la VIFT. Comparaison de deux modèles paramétriques par parties – L'âge à la naissance du premier enfant selon l'union et son type. Estimation pondérée – Femmes âgées de 15 à 80 ans au moment de l'enquête



Nous constatons tout d'abord que les quotients instantanés des deux modèles paramétriques stratifiés sont identiques. Nous constatons ensuite que les taux de la table d'extinction calculés à partir du calcul exact du temps à risque sont identiques aux quotients instantanés des modèles paramétriques stratifiés. On voit donc que le modèle de Poisson, défini en principe pour les événements renouvelables, donne exactement les mêmes résultats que le modèle exponentiel qui est défini pour les événements non renouvelables. On voit également que la table d'extinction, lorsqu'on l'utilise pour opérationnaliser une VIFT par stratification, exprime la comparaison entre des populations théoriques exactement comme le font les modèles statistiques lorsqu'on les utilise en stratifiant selon les modalités de la variable indépendante. On peut présenter ces résultats dans un graphique : le volet de gauche de la figure 1 représente les quotients instantanés obtenus au moyen du modèle de Poisson stratifié.

Le coefficient associé à la modalité « Mariage » des modèles paramétriques où la VIFT est construite de manière stricte signifie bien qu'en moyenne, la probabilité instantanée qu'une femme mariée donne nais-

sance à son premier enfant est 25,6 fois plus élevée que la probabilité qu'une femme qui vit sans conjoint donne naissance à son premier enfant ; le coefficient associé à la modalité « Union de fait » signifie que la probabilité instantanée qu'une femme qui vit en union de fait donne naissance à son premier enfant est 8,6 fois plus élevée que la probabilité qu'une femme qui vit sans conjoint donne naissance à son premier enfant. Le volet de droite de la figure 1 représente le quotient instantané obtenu au moyen du modèle de Poisson où les trois modalités de la situation conjugale sont représentées par une VIFT construite de manière stricte. Dans chaque intervalle, on obtient la valeur du quotient instantané associé au mariage en multipliant la valeur du quotient instantané associé à l'absence d'union dans cet intervalle par 25,6 et on obtient la valeur du quotient instantané associé à l'union de fait en multipliant la valeur du quotient instantané associé à l'absence d'union dans cet intervalle par 8,6. Autrement dit, on applique l'équation $\hat{h}(t) = \hat{h}_0(t)\exp(\beta x)$ qui devient $\hat{h}_M(t) = \hat{h}_{HU}(t) \cdot 25{,}6$ pour le mariage et $\hat{h}_{UL}(t) = \hat{h}_{HU}(t) \cdot 8{,}6$ pour l'union de fait.

Peu importe qu'on les obtienne par stratification ou au moyen d'une VIFT au sens strict, les différences entre les quotients instantanés des différentes modalités de la situation conjugale sont celles qui existent entre les populations théoriques. Dans les deux cas, on présume que les populations théoriques ont généré la population réelle dont on a tiré un échantillon de femmes dont une partie de la biographie, la portion de leur vie vécue au Canada avant d'avoir le premier enfant, a permis de les reconstituer. En comparant les deux volets de la figure 1, nous remarquons que l'usage de la VIFT au sens strict fait disparaître les taux élevés qui, dans les résultats obtenus par stratification, sont associés au mariage entre 15 et 20 ans, ainsi que les taux relativement élevés qui sont associés à l'union de fait au début de la trentaine.

CONCLUSION

Le texte de notre article est tributaire d'un problème de vocabulaire. L'usage veut qu'on distingue nettement la VIFT et la stratification. On enseigne habituellement que la VIFT s'utilise lorsqu'un individu peut se mouvoir d'une modalité à l'autre d'une variable indépendante pendant qu'il est à risque. On enseigne également que la stratification s'utilise lorsque l'effet d'une variable indépendante n'est pas proportionnel, c'est-à-dire lorsque cet effet varie en fonction du temps mesuré depuis l'origine ; on mentionne rarement qu'un individu peut se mouvoir d'une strate à

l'autre pendant qu'il est à risque. Il est rare qu'on aborde la question en posant que l'individu peut se mouvoir d'une modalité d'un caractère à l'autre pendant qu'il est à risque et que la différence entre la VIFT au sens strict et la stratification se ramène simplement au choix d'imposer ou non la proportionnalité, à l'effet de la variable indépendante qu'on construit pour tenir compte de ces mouvements. Nous espérons que notre présentation a éclairci les rapports entre les deux notions.

Cela dit, l'objet principal de l'article est d'expliciter le sens des résultats des modèles statistiques lorsque les individus se meuvent d'une modalité à l'autre d'un caractère pendant qu'ils sont à risque, peu importe que la variable qui sert à représenter ces mouvements soit construite comme une VIFT au sens strict ou au moyen de strates.

Au début de cet article, nous écrivons qu'on lit souvent que les modèles de risque « tiennent compte du passé » ou « possèdent une mémoire », mais que ceci n'est vrai que pour le quotient instantané de base et non pour la variation du quotient en fonction d'une VIFT. Nous expliquons que le calcul du coefficient associé à une VIFT tient compte du temps passé à risque par chaque individu dans chacune des modalités de la VIFT, mais pas de l'ordre dans lequel ces modalités ont été occupées, ni du fait qu'elles aient été occupées successivement ou pas. En termes plus formels, ceci revient à dire que dans les modèles de risque, le quotient instantané de base varie en fonction du temps mesuré depuis l'origine, mais que l'effet de la VIFT est markovien par construction, si on accepte, par analogie, de qualifier de markovien un processus en temps continu : la VIFT ne tient compte que de l'état occupé à chaque instant et pas des états occupés auparavant et l'effet qu'elle mesure ne peut pas être interprété comme un effet conditionnel aux états occupés antérieurement.

Cette idée s'appréhende mieux en comparant la VIFT à la table multirégionale. Tant qu'il demeure dans sa région d'origine, le futur migrant est confondu avec le reste de la population de cette région : on n'admet pas qu'il puisse s'en distinguer par un caractère relié d'une manière ou d'une autre au phénomène qu'on étudie. Plus important, dès le moment de son arrivée dans la région de destination, le migrant est assimilé à la population de cette région avec laquelle son expérience est parfaitement confondue. Ce postulat demeure raisonnable dans la table multirégionale : on peut admettre que le migrant, peu importe ses caractéristiques, ajoutera peu à l'hétérogénéité de la population de la région dans laquelle il s'installe et que l'hétérogénéité qu'il ajoute aura peu d'impact sur la mortalité, ou tout autre phénomène, dans cette région. Surtout, on ne doute pas que les

populations de la région d'origine et de la région de destination du migrant aient existé en tant que populations finies avant sa migration et que chacune continue d'exister ainsi après sa migration. La chose est un peu plus compliquée dans le cas de la VIFT. À quinze ans, peu de femmes sont mariées ou vivent en union de fait ; avant la naissance de leur premier enfant, la plupart auront été mariées ou auront vécu en union de fait. Comme nous l'avons vu, à chaque modalité de la VIFT correspond une population théorique décrite par une série de fonctions. On ne trouve pas de populations finies qui correspondent naturellement à ces populations théoriques au sens où la population finie de chacune des régions d'une table multirégionale existe en tant que population finie en dehors de l'étude de la migration. Les populations théoriques des modalités d'une VIFT n'existent que par le jeu des transitions entre elles, au point où l'unité statistique qui ne passe pas d'une modalité à une autre de la VIFT ne peut même pas être pensée typique de la population théorique qui lui correspond : la femme qui ne se marie pas et ne vit jamais en union de fait entre 15 et 40 ans n'est pas typique de la population théorique de la modalité « Hors union » de notre exemple, pas plus que celle qui est mariée à 15 ans et le demeure jusqu'à la naissance — très rapide dans ce cas — de son premier enfant n'est typique de la population théorique de la modalité « Mariage ». Ce problème existe, peu importe qu'on construise la VIFT sous la forme souple de la stratification ou rigide de la VIFT au sens strict. Par définition, même dans l'étude d'une vraie cohorte, la VIFT construit un jeu de populations théoriques. En utilisant la VIFT, on admet que ce jeu de populations théoriques a engendré la population finie qu'on étudie à partir de l'échantillon qu'on en a tiré. On se trouve à faire une partie de cette opération à l'envers lorsqu'on utilise les quotients instantanés associés à une VIFT, par exemple ceux du tableau 6, pour réaliser une projection par microsimulation : on engendre alors les populations synthétiques de la microsimulation à partir des populations théoriques de la VIFT.

La loi de probabilité qui définit la population théorique associée à chacune des modalités d'une VIFT de même que la comparaison entre les lois des différentes modalités d'une VIFT se conçoit bien lorsqu'on a compris que la VIFT construit un jeu de populations théoriques. Nous avons vu plus haut que le quotient instantané d'une modalité d'une VIFT exprime le risque auquel sont soumises toutes les femmes qui se trouvent dans une situation conjugale donnée et dans une classe d'âge donnée, peu importe l'âge auquel elles sont entrées dans cette situation conjugale. Ainsi, toutes les femmes mariées qui n'ont pas encore donné naissance à

leur premier enfant et qui se trouvent dans un intervalle donné sont soumises au même risque, peu importe qu'elles viennent tout juste de se marier ou qu'elles se soient mariées à 15 ans. De même, la fonction de survie donne la proportion des femmes mariées à 15 ans (*sic*) qui n'auraient toujours pas eu leur premier enfant à la fin de chaque intervalle, même si le quotient instantané, au cours de chacun des intervalles qui suivent le premier, est principalement calculé à partir du temps à risque et des naissances de femmes qui se sont mariées après l'âge de 15 ans. La loi de probabilité associée à chaque modalité de la VIFT résulte de la comparaison entre les populations théoriques associées à chaque modalité ; l'écart entre les populations théoriques est calculé de manière indépendante lorsqu'on construit la VIFT comme une stratification et de manière « moyenne » lorsqu'on utilise une VIFT au sens strict. Peu importe la manière dont elle est construite, le quotient instantané associé à une modalité d'une VIFT ne contient aucune trace du fait d'avoir occupé auparavant une autre modalité de la VIFT et encore moins une mesure de la quantité de temps passé à risque dans cette autre modalité. Plus généralement, la VIFT n'est pas l'opérationnalisation du fait de s'être déplacé d'une modalité à une autre de la VIFT : « Mariage » s'oppose à « Hors union » au sens où en moyenne, le risque des femmes mariées est plus élevé que celui des femmes qui vivent sans conjoint ; la VIFT ne peut pas s'interpréter comme l'opérationnalisation *du fait même* de passer de la modalité « Hors union » à la modalité « Mariage ». En d'autres termes, le risque plus élevé des femmes mariées ne veut pas dire que, pour une femme, le fait de se marier augmente le risque, mais qu'en moyenne, le risque des femmes mariées est plus élevé que celui des femmes qui vivent sans conjoint. Les quotients instantanés associés aux modalités de la VIFT résultent de la comparaison de la survenue des événements aux unités statistiques qui occupent des états distincts à des moments distincts, mais ne résultent pas du fait que ces unités passent d'un état à l'autre même si elles passent en effet d'un état à l'autre.

Il est possible, dans une analyse, de tenir compte du temps passé à risque dans chacune des modalités de la VIFT. Dans notre exemple, on y parviendrait en utilisant une équation un peu plus raffinée dans laquelle, plutôt que de simplement distinguer les trois modalités de la situation conjugale, on conserverait la modalité qui correspond au fait de vivre seule et on utiliserait une série de modalités qui tienne compte du temps écoulé depuis le début du mariage lorsque la femme est mariée, et une autre qui tienne compte du temps écoulé depuis le début de l'union de fait lorsque la

femme est en union de fait. Les modalités d'une telle VIFT pourraient être, par exemple, « Vivre seule », « Être mariée et vivre avec son conjoint depuis moins de trois ans », « Être mariée et vivre avec son conjoint depuis trois à six ans » ou « Être mariée et vivre avec son conjoint depuis au moins six ans », ainsi que trois autres modalités qui correspondraient au même découpage du temps vécu en union de fait. Chaque modalité de cette nouvelle VIFT constituerait une population théorique au sens où, comme nous l'avons expliqué plus haut, chaque modalité d'une VIFT constitue une population théorique distincte.

On peut également construire des équations qui tiennent compte du fait que les individus occupent *successivement* différentes modalités d'un caractère. On réalise ceci en construisant une VIFT dont chacune des modalités correspond à une *trajectoire* différente, c'est-à-dire à une suite d'états occupés successivement. On trouve un exemple de cette manière de faire dans l'article de Martel, Laplante et Bernard (2005) sur les stratégies qu'utilisent les chômeurs et leurs familles pour retrouver le revenu qu'ils avaient avant la perte de leur emploi. Le chômeur qui vient de perdre son emploi peut, entre autres stratégies, choisir de suivre une formation qualifiante pour augmenter ses chances de trouver un emploi ou encore pour augmenter ses chances d'obtenir un emploi meilleur que celui qu'il a perdu. La formation augmente donc le risque de retrouver un emploi ou le risque de retrouver le niveau de revenu perdu. Toutefois, *suivre* la formation qualifiante réduit le risque de se trouver un emploi pendant un certain temps parce qu'on renonce à occuper un emploi pendant qu'on la suit. La formation n'augmente le risque de retrouver un emploi que lorsqu'elle est terminée : *avoir suivi* une formation augmente le risque de trouver un emploi alors que *la suivre* le réduit. Pour appréhender correctement l'effet de la formation, il faut donc distinguer trois modalités d'une seule VIFT — « Ne pas suivre ni avoir suivi de formation », « Suivre une formation » et « Avoir suivi une formation » — dont la dernière contient une trace du passé dans sa définition. Il est donc possible de construire des VIFT qui conservent la trace du passé, en qualité ou en quantité, mais ceci suppose de définir la VIFT et ses modalités dans une véritable conception longitudinale. On n'y parvient pas en conservant simplement les modalités de la définition transversale. La VIFT construite de cette manière n'est pas formellement différente d'une VIFT « ordinaire » : chacune de ses modalités correspond à une population théorique formée d'individus théoriques qui appartiennent à cette population théorique du moment où ils deviennent à risque à celui où ils changent d'état.

Les fonctions de la table d'extinction envisagée comme la réalisation d'une variable aléatoire ressemblent évidemment à certaines fonctions de la table d'extinction en tant qu'outil de la démographie classique ; elles sont évidemment très semblables aux fonctions équivalentes de la démographie mathématique. Au-delà de la forme mathématique, qui est identique, le sens et surtout le contexte sont différents : en statistique mathématique, l'interprétation est résolument probabiliste et inférentielle. La forme mathématique ne représente pas ici un mécanisme qui agit directement *au niveau* de la population ou encore *sur* la population elle-même, la population étant présumée homogène. Elle représente au contraire un mécanisme « moyen » qui agit au niveau d'individus dont on pose, dès le départ, qu'ils sont différents les uns des autres et dont la différence est intrinsèque au modèle (McCullagh et Nelder, 1989). Le rapport entre les deux interprétations de la table d'extinction permet d'envisager la biographie de chaque individu comme élément de la dynamique de la population et la dynamique de la population comme le produit des biographies des individus (Willekens, 2001). C'est en ce sens que la VIFT, qui construit des populations théoriques, se comprend fort mal en dehors de la conception que la statistique mathématique contemporaine offre des rapports entre l'échantillon, la population finie et la population théorique.

BIBLIOGRAPHIE

Binder, D. 1983. « On the variances of asymptotically normal estimators from complex surveys », *International statistical review*, 51 : 279-292.

Binder, D. 1991. « Fitting Cox's proportional hazards models from survey data », *Biometrika*, 79 : 139-147.

Binder, D., et G. Roberts. 2003. « Design-based and Model-based Methods for Estimating Model Parameters », dans R. L. Chambers et C. J. Skinner, dir. *Analysis of Survey Data*. Chichester, UK, Wiley : 29-57.

Chiang, C. L. 1984. *The Life Table and its Applications*. Malabar, FL, R. E. Krieger.

Cox, D. R. 1972. « Regression Models and Life-Tables », *Journal of the Royal Statistical Society, Series B (Methodological)*, 34 : 187-220.

Greville, T. N. E. 1943. « Short methods of constructing life tables », *Record from the American Institute of Actuaries*, 32 : 29-42.

Fisher, R. A. 1922. « On the mathematical foundations of theoretical statistics », *Philosophical Transactions of the Royal Society, A*, 222 : 309-368.

Fisher, R. A. 1931. « The truncated normal distribution », *British Association for the Advancement of Science, Math. Tables*, I : XXXIII-XXXIV.

Hald, A. 1949. « Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point », *Scandinavian Actuarial Journal*, 32 : 119-132.

Hoem, J. M. 1985. « Weighting, misclassification, and other issues in the analysis of survey samples of life histories », dans J. J. Heckman et B. Singer, dir. *Longitudinal Analysis of Labour Market Data.* Cambridge, Cambridge University Press : 249-293.

Kaplan, E. L., et P. Meier. 1958. « Nonparametric estimation from incomplete observations », *Journal of the American Statistical Association*, 53 : 457-481.

Klein, J. P., et M. L. Moeschberger. 2003. *Survival Analysis. Techniques for Censored and Truncated Data, second edition*. New York, Springer.

Korn, E. L., et B. I. Graubard. 1999. *Analysis of Health Surveys*. New York, NY, Wiley.

Leridon, H., et L. Toulemon. 1997. *Démographie*. Paris, Economica.

Lawless, J. F. 1982. *Statistical models and methods for lifetime data*. New York, NY, Wiley.

Martel, E., B. Laplante et P. Bernard. 2005. « Chômage et stratégies des familles, les effets mitigés du passage de l'assurance-chômage à l'assurance emploi », *Recherches sociographiques*, 46 : 245-280.

McCullagh, P., et J. A. Nelder. 1989. *Generalized linear models, second edition*. London, UK, Chapman and Hall.

Morice, E. 1968. *Dictionnaire de statistique*. Paris, Dunod.

Pressat, R. 1973. *L'analyse démographique*. Paris, Presses universitaires de France.

Pressat, R. 1985. *The Dictionary of demography*. Oxford, UK, Blackwell.

Pressat, R. 1995. *Éléments de démographie mathématique*. Paris, Association internationale des démographes de langue française.

Preston, S. H., P. Heuveline et M. Guillot. 2001. *Demography. Measuring and Modeling Population Processes.* Oxford, UK, Blackwell.

Rao, J. N. K., et C. F. J. Wu. 1988. « Resampling inference with complex survey data », *Journal of the American Statistical Association*, 83 : 231-241.

Siegel, J. S. 2002. *Applied Demography*. San Diego, CA, Academic Press.

Willekens, F. J. 2001 « Theoretical and Technical Orientations Toward Longitudinal Research in the Social Sciences », *Canadian Studies in Population* 28 : 189-217.

Wunsch, G., et M. Termote. 1978. *Introduction to demographic analysis. Principles and methods*. New York and London, Plenum Press.

ANNEXE

Variable, caractère, colonne. Depuis l'apparition des progiciels d'analyse statistique, on a pris l'habitude, en sciences sociales, de nommer « variables » les colonnes des fichiers de données (ou des tableaux de nombre) et d'utiliser également le mot « variable » pour désigner chacune des informations qu'on recueille au sujet des unités statistiques au moyen d'un instrument. Cet usage est correct dans la mesure où le mot « variable » est entendu au sens le plus général qu'il peut avoir en science : « symbole ou terme auquel on peut attribuer plusieurs valeurs distinctes, à l'intérieur d'un domaine défini » (Le Petit Robert). Il arrive néanmoins que cet usage prête à confusion. C'est le cas notamment lorsque l'information recueillie auprès des unités statistiques et consignée dans un fichier de données ne correspond pas à la variable dépendante du modèle statistique qu'on estime à partir de cette information. La variable dépendante d'un modèle de risque est le quotient instantané, qui n'est pas mesuré auprès des unités statistiques mais qui est relié à deux informations recueillies auprès des unités statistiques : la quantité de temps pendant laquelle elles ont été à risque et ont été observées, et la manière dont elles ont cessé d'être à risque et observées. C'est également le cas lorsque dans un texte, on traite d'un caractère — c'est-à-dire d'une qualité ou d'une quantité qu'on mesure sur chaque unité statistique — qu'on pense comme une variable aléatoire au sens de la statistique mathématique et qu'on fait de cette variable aléatoire — la « variable dépendante » — une fonction d'autres caractères mesurés auprès de cette unité statistique. Dans cet article, nous discutons un cas où la variable dépendante n'est pas ce qui est mesuré et nous nous concentrons sur la variable dépendante en tant que variable aléatoire. Pour éviter la confusion, nous utilisons le mot « variable » au sens de « variable aléatoire », nous utilisons le mot « caractère » pour désigner l'information recueillie auprès des unités statistiques en tant qu'elle est envisagée du point de vue de la statistique mathématique et le mot « colonne » pour désigner ce à quoi correspond la colonne du fichier de données.

Quotient, taux, risque. Le vocabulaire utilisé en démographie est parfois source de confusion. Pressat (1995 : 1-3) avait proposé de nommer « fonction quotient » ce qui est connu aujourd'hui en français sous le nom de fonction de risque, mais l'usage ne s'est pas imposé. Pressat (1995) de même que Leridon et Toulemon (1997) utilisent indifféremment « quotient instantané » et « taux instantané » pour désigner ce que l'épidémiologie nomme risque instantané, mais Pressat (1973 : 282, 294) semble réserver l'expression « quotient instantané » aux événements non renou-

velables et l'expression « taux instantané » aux événements renouvelables. La démographie de langue anglaise, contrairement à l'épidémiologie, n'a pas de terme spécifique pour le quotient, si bien que, selon les auteurs, il est nommé « *risk* », « *chance* », « *probability* », « *adjusted rate* » et parfois même « *rate* » tout court (*cf.* Wunsch et Termote, 1978 : 19, 22-23). Certains auteurs anglophones posent de manière claire la différence entre la probabilité et le taux (p. ex. Preston, Heuveline et Guillot, 2001 : 1-20), mais d'autres, et pas des moindres, alimentent la confusion en définissant la probabilité comme une forme de taux (p. ex. Siegel, 2002 : 11). La démographie de langue anglaise nomme le risque instantané « *hazard* » et de nombreux auteurs anglophones envisagent cette quantité strictement comme un taux (« *rate* »). Ces usages ajoutent encore à la confusion dans un environnement où le français côtoie l'anglais. On utilisera ici les expressions « quotient instantané » et « fonction de risque ».

ABSTRACT

Benoît LAPLANTE

The nature and meaning of time-varying covariates in demography

It is often said that hazard models "have a memory". Contrary to a common belief, this is true only for the baseline hazard, and not for the effects of time-varying covariates (TVC). The baseline hazard varies according to time measured from the origin, but the effect of a TVC is Markovian by design : the estimation of the effect of a TVC takes into account the state occupied at each moment and not the states occupied previously. Accordingly, its effect is not conditional on previously occupied states. This article looks at this issue from the perspective of mathematical statistics, mainly by distinguishing the *theoretical* population assumed by hazard models and the *real* and *finite* population from which actual samples are drawn. This is done using a simple example : the effect of conjugal status on the first birth, using data from the 2006 *General Social Survey*.