Numéro 2 **Février 2013**

Cahiers
Recherche et Méthodes

Méthodes non-paramétriques de l'analyse des événements du parcours de vie (Event history Analysis)

Estimations avec SPSS
Méthode de Kaplan-Meier et méthode actuarielle

Jean-Marie Le Goff & Yannic Forney

Jean-Philippe Antonietti & André Berchtold Eds.

Université de Lausanne
Faculté des SSP
CH-1015 LAUSANNE

Les Cahiers Recherche et Méthodes (CREM) sont disponibles sur le site web suivant :

http://www.unil.ch/consultation-statistique

Anciens numéros :

1. *Multiple imputation in a longitudinal context: A simulation study using the TREE data*. André Berchtold & Joan-Carles Surís. Janvier 2012.
2. *Méthodes non-paramétriques de l'analyse des événements du parcours de vie (Event history Analysis). Estimations avec SPSS. Méthode de Kaplan-Meier et méthode actuarielle.* Jean-Marie Le Goff & Yannic Forney. Février 2013.

Méthodes non-paramétriques de l'analyse des événements du parcours de vie (Event history Analysis)

Estimations avec SPSS

Méthode de Kaplan-Meier et méthode actuarielle

(Version 2)

Jean-Marie Le Goff

Yannic Forney

Lines
Pôle national de recherché Lives
Université de Lausanne

Méthodes non-paramétriques de l'analyse des événements du parcours de vie (Event history Analysis)

Estimations avec SPSS
Méthode de Kaplan-Meier et méthode actuarielle

Préambule

Le présent cahier a été publié une première fois à la fin 2003 sous le titre « Estimations non-paramétriques avec SPSS. Méthode de Kaplan-Meier et méthode actuarielle » sur une page du site internet du centre Pavie, cette dernière étant consacrée aux méthodes de l'analyse des biographies (*Event history Analysis*). Le site ayant été fermé en 2010, cet article trouve une place naturelle dans la collection des Cahiers de Recherche et Méthodes.
Par rapport à sa version originale, l'article a peu changé. Nous avons fait quelques corrections et ajouté deux références bibliographiques qui nous semblaient importantes. L'article s'appuie sur un usage des procédures *KM* et *LifeTable* de SPSS. Lors de la rédaction de cet article, nous avons utilisé la version 11.5 de SPSS. Si cette version de SPSS est plutôt ancienne, nous avons décidé de maintenir les sorties output que nous avions obtenues avec cette version ainsi que les syntaxes et les images des boîtes de dialogue que nous avons utilisés. En ce qui concerne les syntaxes, celle-ci n'ont pas changé. Les boîtes de dialogue ont parfois légèrement évolué. Les *output* sont quant à eux devenus plus conviviaux et peuvent, en outre, être copiés sur d'autres formats, par exemple, sur un éditeur de texte ou sur un tableau. Néanmoins, ces outputs restent très semblables dans leur design.
Par ailleurs, cet article faisait auparavant partie d'un ensemble de textes consacrés à l'analyse des biographies et les quelques corrections que nous avons apportées ont pour objectif de le rendre autonome. Finalement, le titre a aussi été transformé.
Nous espérons que cette deuxième édition donnera lieu à autant d'interactions entre nous et ses lecteurs qu'il y en a eu lors de sa première publication.

Le 30 janvier 2013
JMLG et YF

Jean-Marie Le Goff est actuellement chercheur au centre de recherche sur les parcours de vie et les inégalités de l'Université de Lausanne et collabore au pole national de recherche LIVES « Surmonter la vulnérabilité : perspective du parcours de vie ». Contacts : jean-marie.Legoff@unil.ch

Yannic Forney est actuellement chef de projet à la Fédération romande des entreprises (Genève)

Sommaire

1. Introduction

Cet article est consacré aux méthodes non-paramétriques d'analyse des événements d'histoire de vie et, plus particulièrement, à la mise en œuvre de ces méthodes avec SPSS. Le terme non-paramétrique signifie ici deux choses :

- Aucune hypothèse n'est posée en ce qui concerne la distribution du risque (*hazard rate*) au cours du temps (Courgeau & Lelièvre 1989, Allison 1995). En d'autres termes, à un instant donné, le risque est estimé de manière totalement indépendante de celui estimé à l'instant précédent ;

- Aucune hypothèse n'est faite en ce qui concerne les différences de rythme d'occurrence des événements au cours du temps entre diverses sous populations.

Les méthodes d'analyse non-paramétriques visent à répondre à un objectif d'exploration des données. Ces explorations consistent à observer la distribution de l'événement étudié au cours du temps, ainsi que de poser ou de tester des hypothèses quant aux différences de distribution entre des sous-populations. Bien que nécessaire, l'utilisation de ces méthodes constitue seulement un préalable à une véritable analyse, pour trois raisons : en premier lieu, l'usage de ces méthodes aboutit à une importante production d'information chiffrée dont il s'agit, souvent à l'aide de graphiques, de tirer les tendances d'évolution des différentes distributions au cours du temps ; en deuxième lieu, on se trouve vite confronté à des problèmes d'effectifs dès lors que l'on veut comparer les distributions de plusieurs sous populations, surtout lorsqu'il s'agit de différencier ces sous-populations en fonction de plus de deux variables ; en troisième lieu, du fait que les individus de ces sous-populations sont loin d'être homogènes du point de vue de leurs caractéristiques, les distributions obtenues de manière non-paramétriques auront peu de similarité, voire aucune, avec celles qui seraient observées dans le cas de populations composées d'individus partageant exactement les mêmes caractéristiques (Vaupel, Menton & Stallard, 1979 ; Vaupel & Yashin, 1985).

Deux types de méthodes d'estimation non-paramétriques de la fonction de survie sont implémentés dans SPSS. La première est la méthode de Kaplan-Meier (KM) et la seconde, la méthode d'estimation actuarielle[1]. Dans la méthode KM, on considère le temps de manière continue. Ainsi l'occurrence d'un événement a lieu strictement à un moment t (Lelièvre & Bringé, 1998). L'usage de cette méthode est particulièrement indiqué lorsque l'on dispose de données portant sur un faible effectif ou pour lesquelles l'unité de temps considérée est petite (par exemple, le mois dans le cas des événements socio-démographiques). En revanche, cette méthode est à déconseiller si de nombreux individus connaissent l'événement à chaque instant ou bien encore si les durées ont été mesurées sur une unité de temps large, par exemple, l'année. Il devient alors plus indiqué d'utiliser les méthodes d'estimation actuarielle. Dans ces dernières, les individus sont agrégés en fonction de l'intervalle de temps au cours duquel ils connaissent l'événement ou sortent d'observation. On considère que les événements et les sorties d'observation se produisent uniformément et indépendamment les uns des autres au cours de l'intervalle de temps compris entre t_i et t_{i+1}. *Ainsi, plutôt que de dire, par exemple,*

[1] Ces deux méthodes sont implémentées dans la plupart des autres package statistique tels que SAS, TDA, etc. Elles semblent ainsi constituer un standard. Dans R, la méthode de Kaplan-Meier est implémentée dans le package Survival alors que la méthode actuarielle peut être estimée à partir du package KMsurv. Il est à noter que les logiciels Stata, S^+ et R offrent une troisième méthode non paramétrique, qui, en l'occurrence, est la méthode de Nelson-Aalen (Allignol, Beyersmann & Schumacher, 2008 ; Bocquier, 1996 ;).

que l'échéance a eu lieu au mois ou à l'année n comme c'est le cas dans la méthode de KM, on dira qu'elle a eu lieu lors du $n^{ième}$ mois ou de la $n^{ième}$ année dans le cas de la méthode actuarielle.

Le deuxième point de cet article sera consacré aux méthodes de Kaplan-Meier alors que la méthode actuarielle sera décrite dans le troisième point. Pour chacune des deux méthodes, nous présenterons succinctement les estimateurs des différentes notions probabilistes de l'analyse des biographies, notamment, la fonction de séjour et le risque cumulé. Dans cette présentation, nous insisterons d'abord sur les principes théoriques d'estimation des distributions avant de développer un exemple d'estimation avec SPSS. Puis, dans chacun des deux cas, notre intérêt portera sur la comparaison des distributions obtenues lorsque l'on différencie plusieurs sous-populations. Dans le quatrième et dernier point, nous ferons quelques remarques de conclusion.

2. Méthode de Kaplan-Meier

2.1 Principe d'estimations

La méthode d'estimation de Kaplan Meier (KM) est aussi appelée par les statisticiens anglo-saxons *Product Limit Estimations* (PLE). Le point central de cette méthode est l'estimation de la distribution de la fonction de séjour *S(t)*, c'est-à-dire, la distribution au cours du temps de la probabilité de ne pas avoir connu l'événement étudié. En d'autres termes, l'intérêt porte plus sur le fait de rester dans une situation que sur la transition vers une autre situation.

2.1.1. Estimation de la fonction de séjour

Si *T*, variable aléatoire, représente la durée écoulée depuis un instant t_0 pour chaque individu avant qu'il n'ait connu l'échéance de l'événement alors :

$$S(t) = P(T > t) \qquad (2.1)$$

Lorsque le temps est considéré de manière discrète, si t_i représente un instant au cours duquel il y a l'observation d'au moins un événement, alors la probabilité de survie au temps t_i est égale à la probabilité d'avoir survécu avant t_i multipliée par la probabilité « conditionnelle » de survivre au temps t_i. L'emploi du terme « conditionnel » veut dire ici qu'il s'agit de la probabilité de survivre au temps t_i *sachant que* les individus étaient survivants en t_i :

$$S(t_i) = S(t_{i-1}) * P(T > t_i / T \geq t_i) \qquad (2.2)$$

Ces différentes probabilités sont estimées à partir des effectifs de population soumis au risque de connaître l'événement, ainsi qu'à partir des effectifs de personnes connaissant les événements en t_i. Il est, en outre, important de souligner que les sorties d'observation, c'est-à-dire, les durées censurées à droite, doivent aussi être prises en compte. Appelons d_i et c_i, les effectifs des individus qui, respectivement, connaissent l'événement et sortent d'observation en t_i. L'effectif N_i des individus soumis au risque de connaître l'événement en t_i correspond à

l'ensemble des individus qui, juste avant que cet instant t_i n'ait été atteint, n'avaient, ni connu l'événement observé, ni n'étaient sortis d'observation[2].

En outre, dans le cas de l'estimation de Kaplan-Meier, on considère que les sorties d'observation c_i ont lieu une fraction de temps après les échéances d_i (Blossfeld & Rohwer, 2001). Dès lors, la proportion h_i des individus qui ont connu l'événement à l'instant t_i correspond à :

$$h_i = \frac{d_i}{N_i} \qquad (2.3)$$

et $(1 - h_i)$ représente la proportion de personnes n'ayant pas connu l'événement. La probabilité de survie en t_i devient alors (cf. équation 2.2) :

$$S(t_i) = S(t_{i-1})(1 - h_i) \qquad (2.4)$$

Par extension, *S(t)* correspond au produit de toutes les probabilités de n'avoir pas connu l'événement depuis le début de l'observation :

$$S(t) = \prod_{t_i \leq t} (1 - h_i) \qquad (2.5)$$

2.1.2 Estimation sur un exemple : durée du premier emploi d'hommes nés entre 1945 et 1964

La mise en œuvre des méthodes non-paramétriques avec SPSS sera tout au long de cet article appliquée sur des données qui concernent la durée du premier emploi d'hommes nés entre 1945 et 1964. Ces données ont été extraites de l'enquête suisse sur la famille réalisée en 1994 (Gabadinho 1998, Gabadinho & Wanner 1999). Cette enquête constitue le volet suisse des *Family et Fertility Surveys* (FFS), c'est-à-dire, un ensemble d'enquêtes réalisées dans 24 pays développés durant la première moitié des années nonante et dont l'objectif était d'analyser les changements dans les modes de constitution de la famille et de la descendance dans ces pays (Macura, Betts & Burkimsher, 2002). En Suisse, ce sont ainsi près de 4000 femmes et 2000 hommes né(e)s entre 1945 et 1975 qui ont été interrogés sur leur biographie professionnelle, migratoire et familiale. En ce qui concerne les activités professionnelles, les personnes interviewées étaient tenues de n'indiquer que les emplois ayant duré au moins trois mois. Notre intérêt porte ici sur la durée du premier emploi de trois mois et plus des hommes nés entre 1945 et 1964. L'événement qui nous intéresse est le départ du premier emploi. En d'autres termes, la transition qui retient notre intérêt se définit comme étant le départ de cet emploi et le début d'une nouvelle situation, qui peut être du chômage, de l'inactivité ou un nouvel emploi. Ceci veut dire ici que la population prise en compte, c'est-à-dire, *la population soumise au risque de connaître l'événement* au début de l'observation est l'ensemble des hommes nés entre 1945 et 1964 qui ont accédé à un premier emploi de trois mois au moins. Ainsi sont écartés de l'analyse les hommes nés après 1964 ou qui n'ont pas accédé à un premier emploi pour une durée de trois mois au moins. L'échantillon est ainsi

[2] En d'autres termes, l'effectif de la population soumis au risque représente l'ensemble des individus qui connaîtront ou qui sortiront d'observation en t_i ou après.

composé de 1 100 personnes ayant eu un premier emploi. Dans cette analyse l'instant initial t_0 est le début de cet emploi et la durée prise en compte est soit la durée qui sépare cet instant initial et le départ de ce premier emploi si celui-ci a été observé, soit la durée qui sépare cet instant initial et le moment de l'enquête si les individus occupaient toujours cet emploi à ce moment-là. L'unité de temps prise en compte est le mois.

Le tableau 2.1 indique l'évolution des différentes grandeurs à prendre en compte dans une estimation de KM durant les premiers et les derniers mois d'emploi. Les colonnes t, N_i, d_i, c_i, h_i, $1\text{-}h_i$ et $S(t)$ représentent respectivement le temps écoulé depuis le début du premier emploi, l'effectif des individus soumis au risque de connaître l'événement (ici, de quitter son premier emploi), l'effectif des individus qui connaissent l'événement, les sorties d'observation, la proportion des individus qui ont connu l'événement en t_i, la proportion des individus n'ayant pas connu l'événement et la fonction de séjour en emploi.

Ainsi, au début du quatrième mois, l'effectif de la population soumise au risque est de 1 100 individus. 54 d'entre eux connaissent l'événement durant ce quatrième mois alors qu'aucun ne sort d'observation. La proportion h_i des individus ayant connu l'événement est de 54/1100, soit 4,91%. La proportion $1\text{-}h_i$ des individus n'ayant pas connu l'événement est de 95,09 %. L'estimateur de $S(t)$ est alors égal à 1*0,9509.

Au début du cinquième mois, la population soumise au risque est de 1046 individus (c'est-à-dire, 1100-54). Les individus qui quittent leur emploi au cours de leur cinquième mois d'activité sont 33 et aucune sortie d'observation n'a lieu pendant ce temps-là. Par conséquent, la proportion d'hommes ayant connu l'événement est de 33/1046, c'est-à-dire 3,15%. Les individus qui n'ont pas connu l'événement durant ce cinquième mois représentent alors 96,85% des individus qui étaient soumis au risque au début du cinquième mois. $S(t)$ correspond au produit de toutes les probabilités de ne pas avoir connu l'événement, c'est-à-dire, 1*0,9509*0,9685 = 0,9209.

Au $309^{ième}$ mois, la population soumise au risque de connaître l'événement n'est plus que de 20 individus. 2 personnes connaissent l'événement et 2 autres sortent d'observation. La proportion d'hommes ayant connu l'événement est de 2/20, c'est-à-dire 10%. Par conséquent, la proportion $(1\text{-}h_i)$ des individus n'ayant pas connu l'événement est de 90%. L'estimateur de $S(t)$ est alors égal à 0,0182*0,9000=0,0164. Ainsi que l'on peut le constater, les événements et les sorties d'observation deviennent fortement aléatoire en fin de période d'observation, c'est-à-dire, lorsque l'effectif des personnes restant soumises au risque de connaître l'événement devient très petit. Le mieux est d'arrêter ces estimations dès lors que cet effectif devient inférieur à 30 individus.

Tableau 2.1 Analyse des départs d'emploi en début et en fin de période d'observation

t_i	N_i	d_i	c_i	h_i	$1-h_i$	$S(t_i)$
3	1100	0	0	0	1	1
4	1100	54	0	0.0491	0.9509	0.9509
5	1046	33	0	0.0315	0.9685	0.9209
6	1013	28	0	0.0276	0.9724	0.8955
7	985	31	0	0.0315	0.9685	0.8673
8	954	34	0	0.0356	0.9644	0.8364
9	920	31	0	0.0337	0.9663	0.8082
10	889	36	0	0.0405	0.9595	0.7755
11	853	20	0	0.0234	0.9766	0.7573
12	833	32	0	0.0384	0.9616	0.7282
13	801	41	0	0.0512	0.9488	0.6909
…	…	…	…	…	…	…
308	21	1	1	0.0476	0.9524	0.0182
309	20	2	2	0.1000	0.9000	0.0164
312	18	2	1	0.1111	0.8889	0.0145
315	16	1	1	0.0625	0.9375	0.0136
320	15	1	1	0.0667	0.9333	0.0127
323	14	1	1	0.0714	0.9286	0.0118
329	13	1	1	0.0769	0.9231	0.0109
330	12	1	1	0.0833	0.9167	0.0100
332	11	2	2	0.1818	0.8182	0.0082
334	9	2	2	0.2222	0.7778	0.0064
344	7	2	2	0.2857	0.7143	0.0045
345	5	1	1	0.2000	0.8000	0.0036
355	4	1	1	0.2500	0.7500	0.0027
357	3	1	1	0.3333	0.6667	0.0018
368	2	1	1	0.5000	0.5000	0.0009
393	1	1	1	1	0	0
404	0	1	1	-	-	-

2.1.3 Variance, écart-type et intervalle de confiance de la fonction de séjour

La méthode de KM étant une méthode *d'estimation* de *S(t)*, il se peut que l'on s'interroge sur l'intervalle de confiance de la probabilité de ne pas avoir connu l'événement à l'instant t_0. Plus généralement *S(t)* étant un estimateur, on peut s'interroger sur sa variance et son écart type. L'estimation de la variance est moins intuitive que ne peut l'être l'estimation de *S(t)*. Nous nous contenterons de donner ici sa formulation, dite formule de Greenwood (Courgeau & Lelièvre, 1989) :

$$\mathrm{var}[S(t)] = [S(t)]^2 \sum_{t_i} \frac{d_i}{N_i(N_i - d_i)} \qquad (2.6)$$

L'écart type de *S(t)* correspondra alors à la racine carrée de la variance de *S(t)* :

$$EC[S(t)] = \sqrt{\mathrm{var}[S(t)]} \qquad (2.7)$$

et l'intervalle de confiance à 95% de *S(t)* (c'est-à-dire un risque d'erreur de l'estimation de *S(t)* fixé à 5%) :

$$IC[S(t)] = S(t) \pm 1{,}96 EC[S(t)] \qquad (2.8)$$

2.1.4 Proportion cumulée des individus ayant connu l'événement et risque cumulé

Quelques autres distributions peuvent être calculées dès lors que l'on connaît la distribution de *S(t).* Il s'agit en premier lieu de la proportion cumulée *F(t)* des individus ayant connu l'événement (en l'absence de sorties d'observation). Cette proportion est le complémentaire à 1 de *S(t)* et va donc indiquer la probabilité d'avoir connu l'événement entre l'instant initial et le moment auquel on se situe :

$$F(t) = 1 - S(t) \qquad (2.9)$$

En deuxième lieu, à partir de *S(t)* peut être estimé le risque cumulé *H(t)* :

$$H(t) = -\log S(t) \qquad (2.10)$$

La distribution de *H(t)* au cours du temps permet alors d'analyser l'évolution du risque au cours du temps (ou de façon plus imagée, la « vitesse » d'occurrence des événements ; cf. Kleinbaum, 1996).

2.2 Estimation de Kaplan-Meier avec SPSS

Dans ce point, nous allons traiter notre exemple des durées d'emploi de trois mois au moins des hommes nés entre 1945 et 1964 au travers de la procédure KM de SPSS[3]. La question de recherche peut être formulée de la manière suivante : comment se distribue au cours du temps la probabilité d'être encore en activité, c'est-à-dire, de ne pas avoir quitté son premier employeur? S'agissant de premiers emplois, on peut supposer qu'un grand nombre d'entre eux auront été occupés pour une courte durée avant que les individus n'aient trouvé, dans un contexte de demande de main-d'œuvre par les employeurs, un autre emploi. On peut néanmoins penser qu'une partie de ces emplois sont déjà des emplois stables dans lesquels les hommes auront tout de suite trouvé des possibilités de carrière interne (Le Goff, 1997). On peut ainsi s'attendre dans un premier temps à une diminution très rapide de la fonction de séjour en emploi, puis dans un deuxième temps à une diminution beaucoup plus lente.

2.2.1. Préparation préalable des données

La préparation des données en vue d'une analyse de Kaplan-Meier est très simple. Tout d'abord, le fichier de données doit être ici un fichier individuel, c'est-à-dire, un fichier composé de l'ensemble des individus soumis au risque de connaître l'événement à l'instant

[3] Nous utilisons ici la version 11.5 de SPSS.

initial[4]. Dans notre exemple, un individu statistique va correspondre à un individu ayant eu au moins un premier emploi de plus de trois mois. Ceci signifie que les hommes n'ayant eu aucun emploi d'une durée d'au moins trois mois ne sont pas pris en compte dans l'analyse.

La mise en œuvre de la procédure KM de SPSS nécessite que les individus soient caractérisés par au moins deux variables :

- *variable de durée*. Cette variable correspond, selon les individus à la variable qui sépare l'instant initial et le moment d'occurrence de l'événement ou le moment de sortie d'observation (censure à droite). Dans notre exemple, cette variable a pour nom « duree » et correspond pour chaque individu à la durée qui sépare le moment d'accès à l'emploi et le moment où il le quitte ou le moment de l'enquête si l'individu n'a pas quitté son premier emploi[5].

  ***Attention** : une précaution doit être prise dans le calcul de la durée lorsque l'on souhaite utiliser la procédure KM de SPSS. En effet, les individus pour lesquels serait associée une durée égale à 0 ne sont pas pris en compte dans l'analyse. Ceci n'est pas le cas dans notre exemple, puisque, par définition, les emplois ont duré au moins trois mois. Ce serait, par contre, le cas si l'on disposait des emplois ayant duré moins de trois mois, notamment, de ceux qui auraient été occupés et quittés un même mois. Par convention, nous considérerons ici que ces durées sont de 1 (que les emplois seraient donc quittés lors du premier mois)[6]. Ceci implique qu'un emploi qui aurait été quitté après un mois d'attente, mais avant la fin du deuxième mois devrait avoir une durée de 2, et ainsi de suite. En conséquence :*

$$durée = ((date\ de\ sortie\ \ d'observation)\ ou\ \ (date\ \ de\ \ l'événement)) - date\ \ de\ \ l'instant\ \ initial + 1$$

  Ce procédé doit être réalisé même s'il n'y a pas de durée initiale égale à 0, c'est-à-dire, lorsque le départ de l'emploi s'effectue le même mois que celui de l'embauche. Ainsi dans notre exemple, les premiers départs d'emploi interviendront à partir du mois 4. Cette méthode peut parfois poser problème. Par exemple, si l'on s'intéresse à la transition de l'école à l'emploi, certaines personnes accéderont directement à une activité professionnelle après leur sortie d'école, sans même un jour d'inactivité ou de recherche d'emploi (Mills, 1999). Or dans ce cas, ils auront une durée d'accès à leur emploi de 1, c'est-à-dire, que l'on considérera qu'il y a eu une période d'inactivité d'un mois ;

- *Indice de censure. Cette variable distingue les individus par au moins deux modalités, celles-ci permettant de distinguer selon que les individus ont connu l'événement*

[4] Il est à noter que dans le cas où on s'intéresserait à la durée de l'ensemble des emplois occupés par les individus au cours de leur carrière, il faudrait en fait créer une ligne pour chaque emploi et disposer d'un fichier « *épisode* » (*spell*).

[5] Dans le cas présent, la seule censure possible est liée au moment de l'enquête. Si, en revanche, l'intérêt portait sur la durée des emplois tant que les individus sont célibataires, il faudrait censurer les individus qui se marient à la date de leur mariage.

[6] Il s'agit d'une convention qui est largement adoptée dans les disciplines des sciences sociales (Blossfeld & Rohwer, 2001). On pourrait toutefois adopter d'autres approches, en considérant, par exemple, qu'un événement a lieu au milieu de l'intervalle de temps (à la moitié du mois considéré) ou au premier quart, etc.

(individus non censurés) ou non (individus censurés). Dans le cas de notre exemple, une variable « censure » à deux modalités a été créée pour chaque individu, la modalité 0 indiquant que les individus occupent toujours leur premier emploi au moment de l'enquête alors que la modalité 1 indique qu'ils ont quitté leur premier emploi. *Cette construction binaire dans laquelle 0 indique qu'il n'y a pas eu observation de l'événement et 1 qu'il a été observé correspond à une convention au sein de la communauté des utilisateurs de l'analyse des biographies*[7]. Une situation plus complexe aurait pu être imaginée, dans laquelle, par exemple, les départs d'emploi auraient été distingués selon qu'ils ont été suivis par une période de chômage, d'inactivité professionnelle, ou par un accès à un deuxième emploi (etc.). La procédure KM de SPSS dans laquelle on déclare la variable de censure est suffisamment souple pour qu'une modalité de départ d'emploi soit considérée comme une échéance ou comme une sortie d'observation, selon l'analyse que l'on veut réaliser.

Si la variable de durée et la variable de censure sont deux variables nécessaires et suffisantes pour procéder à une analyse de KM, d'autres variables se rapportant aux caractéristiques des individus peuvent être prises en compte. Ces variables sont particulièrement intéressantes dès lors que l'on s'intéresse à faire des comparaisons de la distribution de la fonction de séjour entre sous-populations. Nous serons ainsi intéressés plus loin à faire la comparaison de la durée du premier emploi en fonction de la cohorte d'appartenance.

2.2.2 Commandes SPSS

La syntaxe des commandes nécessaires à l'obtention d'une estimation de KM de la distribution des durées d'emploi peut être écrite de la façon suivante :

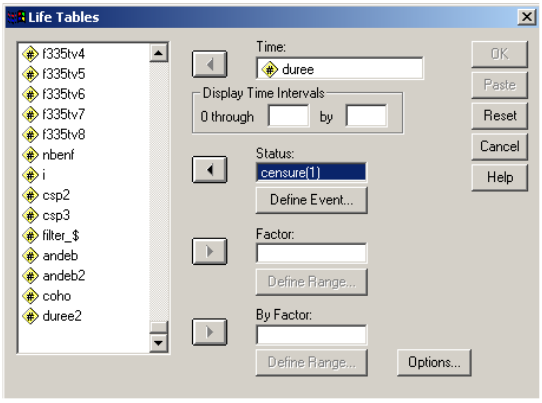
```
KM
duree /STATUS=censure(1)
/PRINT TABLE MEAN
/PLOT SURVIVAL HAZARD.
```

La commande KM indique que l'on souhaite une estimation de Kaplan-Meier. Les autres lignes de programmation permettent de préciser quelles variables seront analysées et ce que l'on souhaite en sortie (output). Ainsi, à la seconde ligne est déclarée la variable de durée, « duree », alors que la commande « STATUS=censure(1) » permet de déclarer le nom de l'indice de censure et la (les) modalité(s) de cette variable correspondant au fait que les individus ont connu l'événement. Ici, est déclaré que l'individu a connu l'événement (dans le cas présent, a quitté son emploi) si la variable « censure » est égale à 1. Toute autre valeur que 1 indique que les individus ont été censurés (dans notre exemple, ils ont tous la valeur 0), c'est-à-dire, n'ont pas connu l'événement. La troisième ligne décrit ce que l'on veut obtenir en « sortie » (output). Il s'agit ici d'une table de survie[8] (qui est désignée par l'option TABLE), ainsi que d'un tableau indiquant la moyenne et la médiane (l'option MEAN). La commande PLOT permet d'avoir aussi en sortie des graphiques, en l'occurrence la représentation de *S(t)* en fonction du temps (SURVIVAL) et de *H(t)* en fonction du temps (HAZARD[9]). Il est à noter que d'autres graphiques peuvent être obtenus[10].

[7] Bien entendu, cela ne veut pas dire que tout le monde suit cette convention.
[8] Cf. description un peu plus loin.
[9] La commande HAZARD est ici mal nommée, en ce sens que le graphique obtenu est la distribution du risque cumulé *H(t)* au cours du temps et non le risque lui même *h(t).*
[10] Cf. plus loin.

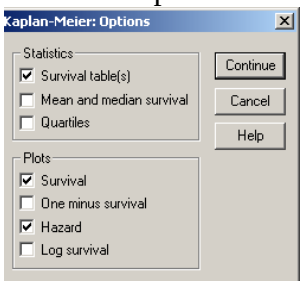
Ces commandes peuvent aussi être définies à travers l'usage des boîtes de dialogue. Ainsi dans le menu *Analyze*, choisir *Survival* puis *Kaplan-Meier*. Définir la durée *(Time)* et l'indice de censure *(status)* qui, ici, est égal à 1 *(Define Event)* dans notre exemple si les individus ont quitté leur emploi (figure 2.1).

Figure 2.1 : Procédure KM par la boîte de dialogue :
définition des variables de durée et de censure



Les options *(options)* permettent d'obtenir la table de survie et les graphiques de *S(t)* et *H(t)* (figure 2.2).

Figure 2.2 : Procédure KM par la boîte de dialogue :
définition des options de sortie



On peut alors « faire tourner » le programme au travers de la commande « *run* ». Néanmoins, l'usage de la fonction coller *(paste)* plutôt que la fonction « *run* » va transcrire les opérations effectuées par l'intermédiaire de la boîte de dialogue en code SPSS sur une page de syntaxe. Ce procédé constitue sans doute la meilleure façon de se familiariser avec les différentes commandes. En outre, disposer de la syntaxe SPSS offre aussi l'avantage d'avoir en mémoire le travail effectué.

2.2.3 Résultats

Les résultats sont édités dans un fichier de sortie (output). En premier lieu est édité une table (tableau 2.2). Dans cette table, chaque ligne correspond à un individu. Les individus sont triés par ordre croissant de durée et de statut (d'abord ceux qui connaissent l'événement, puis ceux qui sortent d'observation). Les colonnes de la table sont :

- 1) *Time* : l'intervalle de temps considéré ($i^{ème}$ mois dans notre exemple) ;

- 2) *Status* : l'indice de status ou de censure c_i;

- 3) *Cumulative Survival* : *S(t)*, la fonction de séjour. Il s'agit donc ici de l'**Estimateur de Kaplan-Meier** ;

- 4) *Standard Error* : l'écart type de *S(t)* ;

- 5) *Cumulative Events* : le cumul des événements (dans notre exemple, le cumul des individus ayant quitté leur emploi) ;

- 6) *Number Remaining* : l'effectif N_i des individus soumis au risque (ici : 1100-événements cumulés avant t_i-sorties totales d'observation avant t_i).

Etant donné la taille de la population que nous avions au départ (N_0=1100 individus), le tableau 2.2 reporte uniquement le début et la fin de cette table. Au début de cette table, on peut retrouver le résultat que nous avions décrit au tableau 2.1, mais de manière agrégée. En ce qui concerne la fin du tableau, au 182^{ième} mois, soit presque 16 ans après avoir accédé à un premier emploi, la probabilité de ne pas avoir quitté cet emploi est de 10% et au 276^{ième} mois (23 ans), de 8%. Ceci montre qu'une part non négligeable des hommes fait carrière chez un unique employeur. La table se termine par un récapitulatif du nombre d'individus présent au début de la période d'observation (*Number of Cases*), du nombre d'individus censurés (*Censored*) ainsi que du nombre d'individus ayant connu l'événement (*Events*) durant toute la période d'observation.

Le tableau 2.3 reporte le tableau de sortie SPSS concernant la durée moyenne de l'emploi et sa durée médiane. Cette moyenne et cette médiane sont accompagnées de leur écart-type et de leur intervalle de confiance à 95%. La moyenne est calculée sur l'ensemble de la période comprise entre l'instant initial t_0 et l'instant maximum au cours duquel est observée une échéance. En d'autres termes, le calcul de cette moyenne ne tient pas compte des sorties d'observation qui auront lieu au-delà de cet instant maximum. La seule consultation de cette moyenne peut donc aboutir à un jugement erroné concernant la distribution de l'événement, surtout si cet événement est rare (il y a à ce moment là de nombreuses sorties d'observation). Dans le cas de données biographiques comportant des sorties d'observation, il est préférable de prendre en compte la durée médiane surtout si l'échantillon comporte de longues durées (Lelièvre et Bringé, 1998). *La médiane représente la durée au cours de laquelle la moitié des individus présents au début de la période d'observation a quitté son emploi* (et donc 50% est encore en activité)[11]. La durée médiane du premier emploi est ici de 25 mois, l'intervalle de confiance à 95% étant compris entre 23,18 et 26,82 mois. En d'autres termes, la moitié des jeunes ont quitté leur premier emploi dans les deux premières années qui ont suivi leur embauche.

Les graphiques de *S(t)* et de *H(t)* sont édités à la suite de ces deux tables et constituent une aide précieuse à l'interprétation. Les croix sur ces graphiques représentent les moments auxquels les individus sortent d'observation (censures). La plupart des sorties d'observations interviennent après le 100^{ième} mois (soit environ 8 ans après l'entrée dans l'emploi). Le graphique de *S(t)* indique que la proportion des personnes se maintenant dans leur premier emploi diminue très rapidement lors des toutes premières années d'observation (Figure 2.3). Néanmoins, cette diminution devient plus faible pour le tiers des individus ayant plus de cinq ans d'activité professionnelle.

[11] Les physiciens accordent une importance particulière à cette grandeur lorsqu'ils s'intéressent, par exemple, à la durée de vie des particules, la médiane prenant alors le nom de « temps de demi-vie ».

> *Le graphique de* H(t) *donne une indication sur l'évolution du risque de connaître l'événement au cours du temps. Plus que le niveau atteint par le risque cumulé, c'est sur l'évolution de la pente de la courbe qu il faut plutôt porter son attention (Courgeau et Lelièvre, 1989). Ainsi, une courbe d'allure convexe indique que le risque diminue au cours du temps. A l'opposé, une pente de forme concave indique une augmentation du risque au cours du temps, alors qu'une droite signifie un risque constant. Dans le cas présent, la courbe de* H(t) *présente une allure générale de type convexe, ce qui signifie que le risque diminue au cours du temps (Figure 2.4).*

Le constat d'une diminution du risque lorsque l'on se situe au niveau d'une population, tel que ceci est observé dans notre exemple, ne peut être extrapolé au niveau individuel. Supposons, par exemple, une société dans laquelle les jeunes femmes sont, soit destinées à se marier, soit destinées à rester célibataire et à occuper une fonction religieuse ; supposons, en outre, que le « risque » de mariage pour les femmes destinées à se marier est constant, quel que soit leur âge (alors qu'il est constamment égal à 0 pour les femmes appartenant à l'autre groupe) ; supposons encore que la distinction entre les deux groupes de femmes se fait en fonction de leur rang de naissance, les aînées étant ainsi celles qui sont destinées au mariage alors que les autres sont destinées aux fonctions religieuses ; supposons, pour terminer, qu'un ethnologue qui observe cette société ne découvre pas cette « variable » de clivage entre les deux sous-populations. Si cet ethnologue est aussi un statisticien féru d'analyse des biographies, il sera à n'en pas douter intéressé à estimer la distribution de *S(t)* (probabilité de ne pas être mariée) et de *H(t)* (risque cumulé de mariage) en prenant au départ l'ensemble de la population des femmes célibataires, qu'elles soient aînées ou non. Au fur et à mesure de l'écoulement du temps, les femmes aînées vont se marier et prendre de moins en moins de poids dans la population des femmes soumises au risque de se marier. Par conséquent, la population des femmes destinées à rester célibataires va prendre de plus en plus de poids et le risque ira dans le sens d'une diminution au cours du temps. La courbe de *H(t)* obtenue par l'ethnologue sera d'allure convexe, alors que le tracé de la courbe de *H(t)* dans le cas où seules les femmes aînées seraient prises en compte serait une droite. Cet exemple fictif montre que la variation du risque au cours du temps peut dépendre de la composition de la population et de son hétérogénéité, même si cette hétérogénéité n'est pas observée (Vaupel & Yachin, 1985). Une interprétation erronée de la part de notre ethnologue-statisticien serait de conclure que plus la situation de célibat s'allonge, plus les « chances » de mariage diminuent. De la même manière, si l'on revient à notre exemple, réel celui-ci, de la durée du premier emploi en Suisse, *le constat d'une diminution du risque de départ de l'activité professionnelle au cours du temps ne permet pas de conclure que plus la durée de l'emploi s'allonge, plus le risque pour un individu de quitter cet emploi diminue.*

Un regard plus attentif sur l'évolution de la pente de *H(t)* peut être porté, en considérant l'hypothèse selon laquelle cette évolution dépend de l'hétérogénéité de la population (Figure 2.4). Il semble que la courbe soit quasiment droite durant les cinq premières années d'observation, avant de diminuer de plus en plus fortement. Une telle figure semble ainsi indiquer l'existence de deux sous-populations, la première se caractérisant par un risque important de départ d'activité, la seconde par un risque plus faible, cette seconde sous-population prenant de plus en plus de poids au fur et à mesure que les individus de la première quittent leur activité professionnelle. Un tel schéma apparaît ainsi conforter l'hypothèse que nous avions faite en amont de la mise en œuvre de la procédure de Kaplan-Meier et selon laquelle si un premier emploi est pour un nombre important de jeunes un emploi de passage, il peut aussi être un emploi d'insertion pour un nombre non négligeable d'autres jeunes.

Tableau 2.2 : Table SPSS des résultats de l'analyse de Kaplan Meier sur le premier emploi des hommes nés entre 1945 et 1964.

Début de la table

Survival Analysis for DUREE

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
4.00	1.00			1	1099
4.00	1.00			2	1098
4.00	1.00			3	1097
4.00	1.00			4	1096
4.00	1.00			5	1095
4.00	1.00			6	1094
4.00	1.00			7	1093
4.00	1.00			8	1092
4.00	1.00			9	1091
4.00	1.00			10	1090
4.00	1.00			11	1089
4.00	1.00			12	1088
4.00	1.00			13	1087
4.00	1.00			14	1086
4.00	1.00			15	1085
4.00	1.00			16	1084
4.00	1.00			17	1083
4.00	1.00			18	1082
4.00	1.00			19	1081
4.00	1.00			20	1080
4.00	1.00			21	1079
4.00	1.00			22	1078
4.00	1.00			23	1077
4.00	1.00			24	1076
4.00	1.00			25	1075
4.00	1.00			26	1074
4.00	1.00			27	1073
4.00	1.00			28	1072
4.00	1.00			29	1071
4.00	1.00			30	1070
4.00	1.00			31	1069
4.00	1.00			32	1068
4.00	1.00			33	1067
4.00	1.00			34	1066
4.00	1.00			35	1065
4.00	1.00			36	1064
4.00	1.00			37	1063
4.00	1.00			38	1062
4.00	1.00			39	1061
4.00	1.00			40	1060
4.00	1.00			41	1059
4.00	1.00			42	1058
4.00	1.00			43	1057
4.00	1.00			44	1056
4.00	1.00			45	1055
4.00	1.00			46	1054
4.00	1.00			47	1053
4.00	1.00			48	1052
4.00	1.00			49	1051
4.00	1.00			50	1050
4.00	1.00			51	1049
4.00	1.00			52	1048
4.00	1.00			53	1047
4.00	1.00	.9509	.0065	54	1046

Fin de la table

240.00	.00			988	45
248.00	.00			988	44
249.00	1.00	.0892	.0092	989	43
251.00	.00			989	42
256.00	1.00	.0871	.0092	990	41
256.00	.00			990	40
257.00	.00			990	39
257.00	.00			990	38
263.00	.00			990	37
265.00	.00			990	36
267.00	.00			990	35
271.00	.00			990	34
271.00	.00			990	33
271.00	.00			990	32
272.00	.00			990	31
275.00	.00			990	30
276.00	1.00	.0842	.0094	991	29
276.00	.00			991	28
276.00	.00			991	27
280.00	.00			991	26
283.00	.00			991	25
287.00	.00			991	24
298.00	.00			991	23
299.00	.00			991	22
308.00	.00			991	21
309.00	.00			991	20
309.00	.00			991	19
312.00	.00			991	18
312.00	.00			991	17
315.00	.00			991	16
320.00	.00			991	15
323.00	.00			991	14
329.00	.00			991	13
330.00	.00			991	12
332.00	.00			991	11
332.00	.00			991	10
334.00	.00			991	9
334.00	.00			991	8
344.00	.00			991	7
344.00	.00			991	6
345.00	.00			991	5
355.00	.00			991	4
357.00	.00			991	3
368.00	.00			991	2
393.00	.00			991	1
404.00	.00			991	0

Number of Cases: 1100 Censored: 109 (9.91%) Events: 991

Tableau 2.3 : Durée moyenne et médiane des emplois (Résultats SPSS)

	Survival Time	Standard Error	95% Confidence Interval	
Mean:	67.23	3.40	(60.56,	73.90)
(Limited to	404.00)			
Median:	25.00	.93	(23.18,	26.82)

Figure 2.3 : Fonction de séjour en emploi (estimation de Kaplan-Meier)

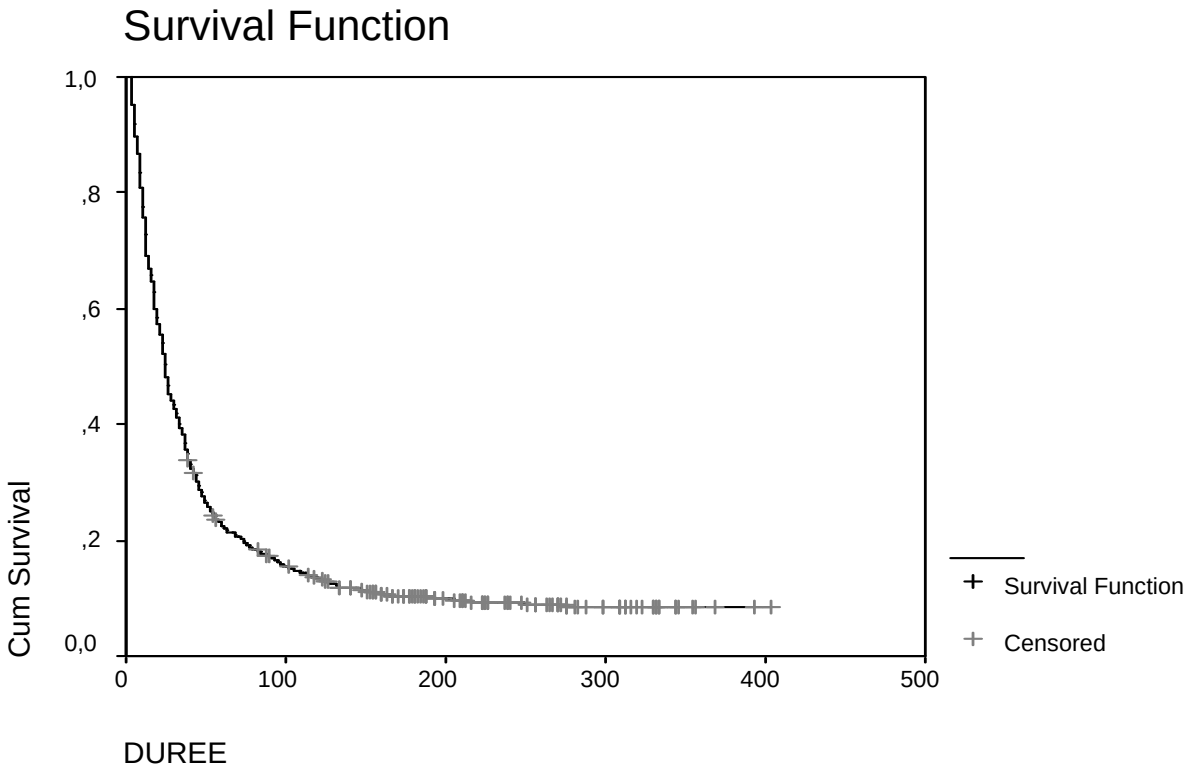
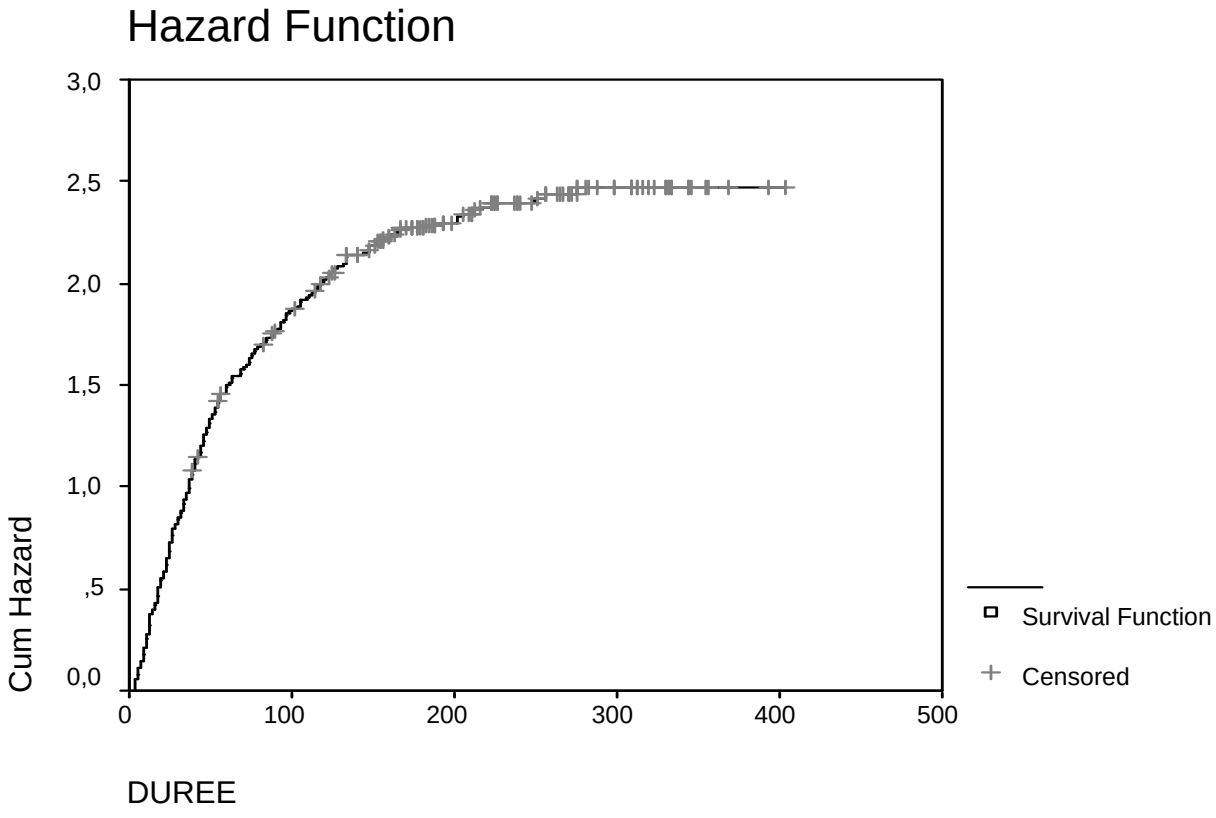


Figure 2.4 : Fonction de risque cumulé de départ d'emploi (estimation de Kaplan-Meier)



2.3. Comparaison des distributions entre plusieurs sous-populations

Quelles sont les différences dans la durée du premier emploi selon le moment de naissance ? Dans le cas présent, il est important de rappeler que les hommes pris en compte sont nés entre 1945 et 1964. En conséquence, le moment de l'insertion professionnelle s'étale entre 1960 et le milieu des années nonante dans cette population. Dans le cas des années soixante, voire septante, l'insertion professionnelle a lieu dans un contexte dans lequel la stabilisation dans l'emploi est valorisée et constitue une norme. En revanche, à partir des années quatre-vingt, se diffusent de nouvelles techniques de management dans lesquelles, notamment, est émise l'idée que les employés doivent être mobiles et qu'une carrière doit se réaliser au sein de plusieurs entreprises et non plus dans une seule (Boltanski & Chiappello, 1999). A cet aspect de changement dans les normes associées à la notion de carrière professionnelle peut, en outre, parfois s'ajouter des difficultés à se maintenir en emploi pour des raisons se rapportant aux problèmes économiques que connaissent beaucoup d'entreprises dans le sillage des chocs pétroliers des années septante et quatre-vingt.

2.3.1 Deux sous-populations

Notre hypothèse de différenciation de la durée du premier emploi peut être testée au travers d'une comparaison des distributions des fonctions de séjour selon la cohorte de naissance. Dans le cas présent, la population va être subdivisée en deux, les hommes nés entre 1945 et 1954 d'une part, ceux nés entre 1955 et 1964 d'autre part. Nous allons ici utiliser des tests statistiques permettant de comparer les distributions des échéances au cours du temps.

a) Tests de comparaison implémentés dans SPSS

Dans SPSS, les tests qui peuvent être mis en œuvre en vue de comparer deux ou plusieurs sous-populations sont ceux dits du *Log-Rank*, de *Breslow* et de *Tarone-Ware*. Ces trois tests statistiques sont très voisins les uns des autres et sont apparentés aux tests de rang. L'idée générale consiste à comparer la distribution d'occurrence des événements dans chacune des sous-populations avec l'hypothèse nulle H_o que les distributions des fonctions de séjour sont semblables dans chacune des sous-populations. Chaque test statistique s'appuie sur le calcul d'une grandeur U qui correspond au temps t_i de la somme de chaque écart entre le nombre d'échéances observées (O_i) et le nombre d'échéances attendues (E_i), multiplié par une pondération w_i :

$$U = \sum_i w_i (0_i - E_i) \quad (2.11)$$

Le nombre E_i correspond au nombre d'échéances qui seraient observées au temps t_i si les deux distributions étaient strictement identiques Les trois tests diffèrent entre eux par la pondération w_i qui est prise en compte. Dans le cas du test du Log-Rank, tous les poids sont égaux à 1, ce qui revient à dire que le même poids est donné à chaque événement ; pour le test de Breslow, le poids correspond à la population soumise au risque en t_i ($w_i=n_i$) et à sa racine carrée ($w_i=\sqrt{n_i}$) dans le cas du test de Tarone Ware (Blossfeld & Rohwer, 2001). Dans le cas d'une comparaison entre deux sous-populations, les trois statistiques obtenues doivent être comparées à un χ^2 de 1 degré de liberté.

Il est important de mentionner que les tests de Breslow et de Tarone-Ware donnent plus de poids aux événements qui ont eu lieu en début de période d'observation, lorsque la population

soumise au risque est encore grande, par rapport à ceux ayant lieu en fin de période d'observation où le poids donné aux événements devient plus faible. *En conséquence, les deux tests de Breslow et de Tarone Ware sont plus sensibles à des différences dans les distributions d'occurrence des événements en début de période d'observation. En revanche, le test du Log-Rank est plus sensible à des différences existantes en fin de période d'observation (Blossfeld & Rohwer, 2001).*

b) Syntaxe SPSS

Les options de commande de la procédure de KM permettent d'estimer les fonctions de séjour de différentes sous-populations et de faire appel aux trois tests statistiques que nous venons de mentionner. Avant même la mise en œuvre de cette procédure de Kaplan Meier, nous avons au préalable défini une variable dichotomique « *coho* » qui est égale à 1 si les individus sont nés entre 1945 et 1954 et égale à 2 s'ils sont nés entre 1955 et 1964. La syntaxe KM a été écrite de la façon suivante :

```
KM
  duree BY coho /STATUS=censure(1)
  /PRINT MEAN
  /PLOT SURVIVAL HAZARD
  /TEST LOGRANK BRESLOW TARONE
```

La procédure de Kaplan-Meier est, comme précédemment, appelée dans la première ligne de programmation. La seconde ligne permet toujours de déclarer la variable de durée « duree ». Toutefois, en ajoutant la syntaxe « BY coho », on déclare ici que l'on souhaite obtenir une estimation de KM pour chacune des deux cohortes décennales. « STATUS= censure (1) » reste inchangé. Dans l'option PRINT, nous n'avons pas repris la mention « TABLE » : en effet, ainsi que nous l'avons vu dans le point précédent, la table de survie donnée en sortie est très longue et donne une trop grande quantité d'information chiffrée qu'il est facile à résumer ou à traduire à l'aide de graphiques[12]. Ces graphiques sont demandés avec l'option PLOT, de la même manière que précédemment. Dans ce cas, les distributions de *S(t)* pour chacune des deux sous-populations sont représentées sur un même graphique, de même en ce qui concerne les distributions de *H(t)*. A la cinquième ligne, les tests du Log-Rank, de Breslow et de Tarone-Ware sont appelés avec l'option TEST.

Comme précédemment, ces commandes peuvent être appelées avec les boîtes de dialogue en sélectionnant dans le menu *Analyse*, *Survival* puis *Kaplan-Meier*. Choisir ensuite la durée (*Time*) et l'indice de censure (*status*), puis définir la variable (*Factor*) qui permet de décomposer la population de départ en plusieurs sous-populations (ici *coho*). Ensuite, cliquer sur *Compare Factor* pour sélectionner chacun des trois tests statistiques.

[12] Dans le cas présent, la mention de cette option se traduirait par l'édition de deux tables, une pour chaque cohorte.

c) Résultats

Les tests statistiques et les graphiques sont édités à la suite des tables dans un fichier de sortie (output). Le tableau 2.4 donne les résultats des tests, les degrés de liberté (*df*) et la significativité du test.

Le χ^2 calculé pour le test du Log-Rank est égal à 4,53 et est à comparer au χ^2 théorique de 3,84 pour un risque α=5%. Il en est de même pour les deux autres tests. Dans chaque cas, le χ^2 observé est supérieur au χ^2 théorique, ce qui nous conduit à rejeter l'hypothèse statistique d'homogénéité des distributions de durée. En d'autres termes, la distribution de la durée des emplois diffère entre les deux cohortes et ce, aussi bien en début qu'en fin de période d'observation. Dans les tables SPSS, les χ^2 sont accompagnés de leur *p-valeur* (colonne *« significance »* dans le tableau 2.4). Une *p-valeur* inférieure à α = 5% indique que l'hypothèse nulle doit être rejetée. Notre hypothèse selon laquelle la durée du premier emploi est plus longue chez les hommes nés entre 1945 et 1954 que chez ceux nés entre 1955 et 1964 apparaît ainsi se vérifier.

Les graphiques 2.5 et 2.6 représentent respectivement la fonction de séjour *S(t)* et la fonction de risque cumulé *H(t)* pour chacune des deux cohortes. *S(t)* diminue rapidement dans chacune des deux sous-populations en début de période. Des différences entre les deux cohortes apparaissent clairement après 50 mois d'activité professionnelle (environ 4 ans). De même, la fonction *H(t)* est relativement semblable pour les deux sous-populations en début de période, puis les cohortes connaissent des évolutions particulières à partir du cinquantième mois. Le risque de quitter son emploi diminue plus rapidement pour la cohorte 1945-1954 que pour la cohorte 1955-1964.

Tableau 2.4 : Résultats des tests de comparaison entre deux sous-populations

Test Statistics for Equality of Survival Distributions for COHO

	Statistic	df	Significance
Log Rank	4.53	1	.0333
Breslow	4.15	1	.0417
Tarone-Ware	4.32	1	.0377

Figure 2.5: Fonction de séjour en emploi selon la cohorte de naissance (estimation de Kaplan-Meier)

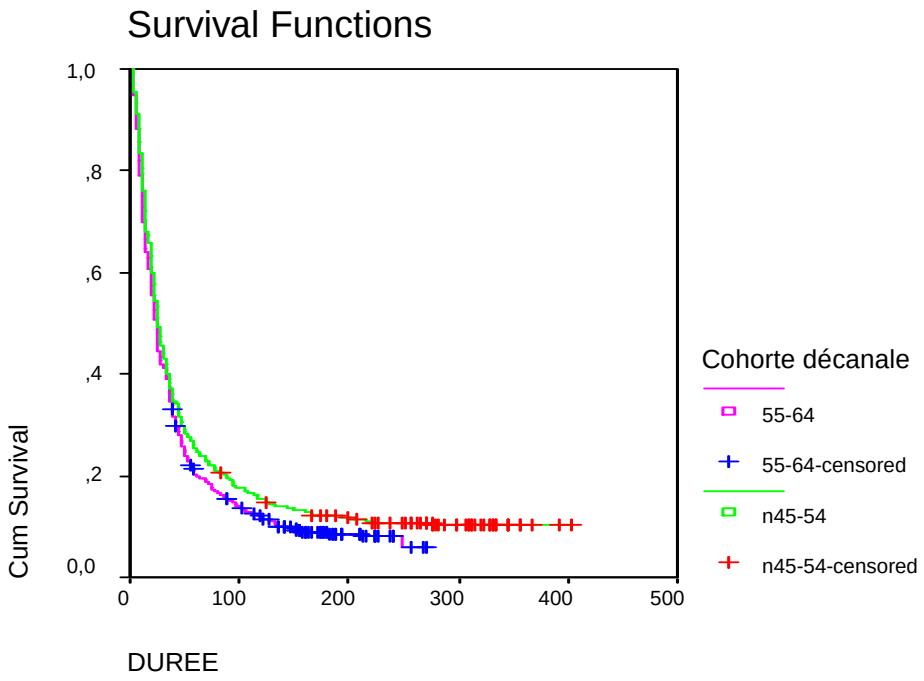
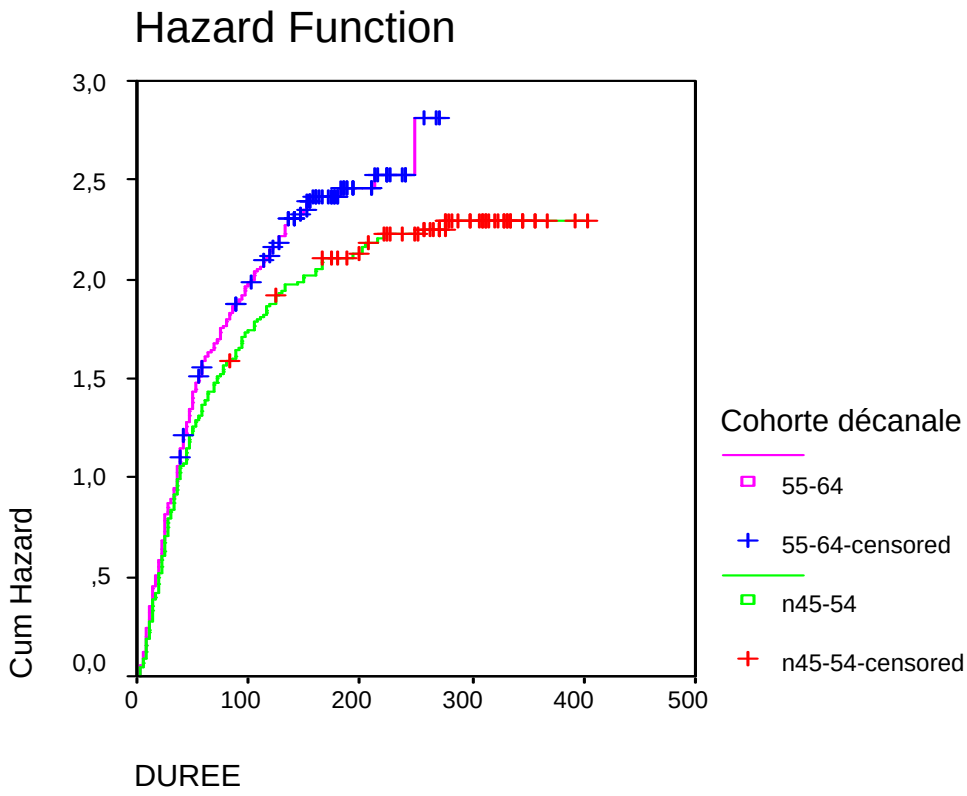


Figure 2.6: Fonction de risque cumulé de départ d'emploi selon la cohorte de naissance (estimation de Kaplan-Meier)



2.3.2 Plus de deux sous-populations

a) Syntaxe SPSS

Notre intérêt porte maintenant non plus sur l'analyse de deux cohortes décennales, mais de quatre cohortes quinquennales définies par une variable du nom « *akl5* » dans laquelle les modalités 6, 5, 4 et 3 indiquent que les individus sont nés respectivement entre 1945 et 49, 1950 et 54, 1955 et 59, 1960 et 64[13]. Par commodité, nous appellerons ces modalités, *facteurs*, afin de correspondre au terme *factor* tel que celui-ci est utilisé dans la procédure KM de SPSS. Les commandes suivantes ont été écrites de la même manière que précédemment à l'exception du fait que « *akl5* » remplace « *coho* » dans la deuxième ligne :

```
KM
  duree BY akl5 /STATUS=censure(1)
  /PRINT MEAN
  /PLOT SURVIVAL HAZARD
  /TEST LOGRANK BRESLOW TARONE
  /COMPARE OVERALL POOLED
```

La dernière ligne « COMPARE OVERALL POOLED » signifie que les tests vont comparer les écarts de distribution de l'une ou l'autre des sous-populations à la distribution qui serait observée si l'on disposait de l'ensemble de la population. Nous montrerons aussi les résultats de l'option :

```
/COMPARE PAIRWISE, STRATA
```

Cette dernière commande permet de comparer chacune des sous-populations deux à deux. Ainsi, s'il y a trois sous-populations, il y aura trois séries de tests, alors que s'il y a quatre sous-populations, il y aura six séries de tests.

b) Résultats

Les résultats des trois tests statistiques dans le cas de la comparaison de l'ensemble des sous-populations sont présentés dans le tableau 2.5. Le nombre de degré de liberté (*df*) correspond au nombre de sous-populations moins une. Comme ici quatre sous-populations sont prises en compte, le nombre de degrés de liberté est de 3. Le χ^2 calculé pour le test du Log-Rank est égal à 4,58 et est par conséquent plus petit que le χ^2 théorique de 7,81 pour un risque α=5%. Ce résultat est également valable pour les deux autres tests et montre que l'hypothèse nulle ne peut être rejetée. Ceci signifie que les distributions de la durée des emplois ne diffèrent pas entre les sous-populations.

Un tel résultat semble en contradiction avec le résultat précédent dans lequel la subdivision de l'ensemble de la population en seulement deux sous-populations donnait un résultat significatif. Cette contradiction semble toutefois devoir être liée avec le fait que la subdivision en quatre sous-populations mène à des effectifs faibles, et en conséquence, les tests statistiques deviennent moins sensibles à des différences entre cohortes.

Dans le cas de l'option qui permet de comparer les populations deux à deux, puisque nous disposons de quatre sous-populations, six séries statistiques sont réalisées. Le tableau 2.6

[13] Il s'agit en fait d'une variable faisant partie du fichier original de l'enquête sur la famille dans lequel les modalités 1 et 2 étaient associées respectivement aux individus nés entre 1970 et 1975 et 1965 et 1969.

reporte les résultats de ces tests. Par exemple, la comparaison des cohortes 1960-64 et 1955-59 (facteurs 3 et 4) dans le cas du test du Log-Rank a pour résultat 0,8373, valeur non significative pour un seuil fixé à 5%. En général, les valeurs montrent que les trois séries de tests statistiques ne sont pas significatifs puisque les *p-valeurs* sont supérieures à 5%.

Tableau 2.5 : Résultats des tests dans le cas de plus de deux sous-populations (option « /compare overall pooled »).

Test Statistics for Equality of Survival Distributions for AKL5

	Statistic	df	Significance
Log Rank	4.58	3	.2053
Breslow	4.58	3	.2051
Tarone-Ware	4.50	3	.2127

Tableau 2.6 : Résultats des tests dans le cas de plus de deux sous populations (option « compare pairwise, strata »)

Log Rank Statistic and (Significance)

Factor	3	4	5
4	,04 (,8373)		
5	1,99 (,1583)	2,69 (,1009)	
6	1,65 (,1988)	2,61 (,1061)	,00 (,9583)

Breslow Statistic and (Significance)

Factor	3	4	5
4	,21 (,6441)		
5	2,24 (,1342)	3,74 (,0532)	
6	,80 (,3726)	1,79 (,1805)	,20 (,6555)

Tarone-Ware Statistic and (Significance)

Factor	3	4	5
4	,10 (,7480)		
5	2,19 (,1385)	3,25 (,0716)	
6	1,16 (,2809)	2,07 (,1499)	,05 (,8180)

2.4. Précisions sur la commande KM de SPSS

Outre les instructions que nous avons déjà décrites, celles qui suivent peuvent aussi être mises en œuvre en vue de spécifier les résultats que l'on souhaite obtenir en fichier de sortie :

/PRINT =TABLE, MEAN, NONE

Par défaut (si PRINT est omis), KM donne la table de survie et le tableau de la moyenne et de la médiane avec leur écart type et intervalle de confiance. Si PRINT est spécifié, alors des mots-clés comme TABLE, MEAN ou NONE peuvent être ajoutés (SPSS, 1993). PRINT TABLE permet d'avoir en sortie la table de survie de Kaplan Meier (une pour chaque sous-population) alors que l'instruction MEAN permet d'avoir un tableau récapitulant la moyenne, la médiane ainsi que les écarts types et les intervalles de confiance de ces deux grandeurs. PRINT NONE correspond au cas où l'on ne souhaite pas de tables en sortie, ce qui peut être utile si l'on désire disposer seulement des tests ou des graphiques.

/PLOT = ALL, SURVIVAL, HAZARD, OMS, LOGSURV, NONE

L'absence d'une instruction PLOT conduit à n'obtenir aucun graphique dans le fichier de sortie. Il en est de même lorsque l'option /PLOT NONE est spécifiée. Si la commande /PLOT est écrite sans spécification, alors sera produit dans le fichier de sortie uniquement un graphique de la fonction de séjour au cours du temps. Le mot-clé HAZARD permet d'obtenir le graphique du risque cumulé[14] de connaître l'événement au cours du temps, alors que la mention OMS (*One Minus Survival*)[15] permet d'avoir en sortie un graphique de la distribution de la proportion cumulée *F(t)* des individus ayant connu l'événement considéré. Le mot-clé LOGSURV permet d'avoir un graphique de la distribution de *S(t)* sur une échelle logarithmique. L'option PLOT ALL permet d'avoir en sortie tous les graphiques.

/PERCENTILES

Par défaut, cette instruction donne les quartiles pour chaque combinaison de facteurs ou de strates, c'est-à-dire, les durées au cours desquelles 25%, 50% (médiane) et 75% (s'il y a lieu) des personnes ont connu l'événement. Le résultat apparaît dans le fichier output entre la table de survie et les graphiques. Toutefois, la commande /PERCENTILES = (x_1 x_2 etc.) permet d'obtenir en sortie les durées pour lesquelles x_1, x_2% de la population a connu l'événement (x_1, x_2, etc. doivent être compris entre 0 et 100).

/SAVE= SURVIVAL SE HAZARD CUMEVENT

Cette instruction permet de garder dans le fichier de données (data) des variables temporaires crées par KM (SPSS, 1993). Il s'agit de : SURVIVAL (fonction de survie) ; SE (écart type de *S(t)*) ; HAZARD qui est ici le risque cumulé *H(t*) et non *h(t)* ; CUMEVENT (Nombre d'événements cumulés). Il est à noter que les individus qui ne connaissent pas l'événement apparaissent dans le fichier comme des données manquantes.

[14] Rappelons l'imprécision de SPSS concernant ce graphique. Il s'agit bien d'un graphique de *H(t*) et non de *h(t*). *H(t)* est spécifié ici par l'opposé du logarithme à chaque instant de *S(t).*

[15] C'est-à-dire *1-S(t).*

/STRATA= nom de variable

Cette instruction permet de stratifier une variable donnée. Cependant, l'analyse s'effectue à un niveau supérieur par rapport à l'instruction « BY ». Par exemple, si nous avions disposé de la population des femmes en plus de celle des hommes et que nous souhaitons stratifier la variable sexe (1 pour les hommes et 2 pour les femmes), alors la commande KM aurait été écrite de la façon suivante :

```
KM
  duree BY coho /STATUS=censure(1)
  /STRATA sexe
  /PRINT TABLE
  /PLOT SURVIVAL HAZARD.
```

L'instruction STRATA correspond ici à une répartition par sexe de la durée du premier emploi en tenant compte à chaque fois des différences entre cohortes. L'option « /STRATA » se différencie de l'option « BY » en ceci que dans cet exemple nous obtiendrions deux graphiques, un pour chaque sexe, chaque graphique reportant les distributions de chacune des deux cohortes. En outre, il n'y a pas de test de comparaison avec l'option « /STRATA ».

/TREND

Cette commande permet de tenir compte des écarts de distribution existant entre différentes sous-populations (définies par « BY variable »). Dans le cas présent, par défaut l'écart entre chacune des quatre sous-populations sera le même, c'est-à-dire, que l'on considère que la différence (écart) entre les femmes nées entre 1945 et 1949 et celles nées entre 1950 et 1954 sera la même que celle que l'on observe pour les femmes nées entre 1950 et 1954 et celles nées entre 1955 et 1959, ainsi qu'entre les femmes nées entre 1955 et 1959 et celles nées entre 1955 et 1959 et celles nées entre 1960 et 1964. Toutefois, on peut supposer que l'écart entre les cohortes 1950-54 et 1955-59 sera plus important que celui qui sépare d'une part les cohortes 1945-49 et 1950-54 d'une part, et les cohortes 1955-59 et 1960-64 d'autre part. Ceci nous est suggéré par le fait que les tests étaient significatifs lorsque la population était distinguée en deux cohortes décennales alors qu'ils ne l'étaient plus lorsqu'elle était distinguée en quatre cohortes quinquennales.

L'option « /TREND (-3 -1 1 3) » sera ici la métrique par défaut, c'est-à-dire, que les écarts seront considérée semblables entre les cohortes[16]. Dans le cas présent, l'usage de cette métrique indique aboutit à des tests qui ne sont pas significatifs (non montrés ici). En revanche, pour tester l'hypothèse d'un plus grand écart entre les deux cohortes intermédiaires, l'option peut être écrite ainsi : « /TREND (-5 -3 3 5) »[17]. Il est à noter que du fait que l'on fait des hypothèses sur les écarts, le nombre de degré de liberté des tests de comparaison ne sera plus de trois (nombre de sous-population moins une) mais de un. Dans le cas présent, le résultat du test du *Log-Rank* est égal à 4,13 (tableau 2.7) et est à comparer avec un χ^2 théorique de 3,84. Un tel résultat semble bien confirmer l'idée qu'il y a une différence de distribution dans les durées d'emploi entre les cohortes 50-54 et 55-59, mais pas de différences entre les cohortes 1945-49 et 50-54 d'une part, et entre les cohortes 1955-59 et 60-64 d'autre part. Néanmoins, les autres tests apparaissent significatifs seulement au seuil de 10%, ce qui doit nous inciter à prendre ce résultat avec prudence.

[16] Dans le cas où on a trois sous-populations et non quatre, on écrit /TREND (-1 0 1) : il y a toujours le même écart entre les chiffres dans les parenthèses.

[17] L'écart entre les deux chiffres intermédiaires est ici plus élevé (il est de 6) qu'entre les autres (il est de 2).

Tableau 2.7 : Résultats des tests dans le cas de plus de deux sous-populations (option « /trend »)

Test Statistics for Equality of Survival Distributions for AKL5
with Trend, metric = (-5, -3, 3, 5)

	Statistic	df	Significance
Log Rank	4.13	1	.0421
Breslow	3.28	1	.0702
Tarone-Ware	3.65	1	.0559

/ID = varname

L'instruction ID permet de donner un label à la variable (ici *coho*). En d'autres termes, si notre ID est la variable cohorte (exemple ci-dessous), alors dans notre fichier output se rajoutera une colonne avec le label des différentes cohortes.

ID	Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
n45-54	4,00	1,00			1	1099
n45-54	4,00	1,00			2	1098
n45-54	4,00	1,00			3	1097
n45-54	4,00	1,00			4	1096
etc.						
55-64	4,00	1,00			23	1077
55-64	4,00	1,00			24	1076
etc.						

*** * ***

Dans ce point, nous avons vu comment utiliser les méthodes d'estimation de Kaplan-Meier avec SPSS. Rappelons ce que nous avions souligné dans l'introduction de cet article, à savoir que l'usage de ce type de méthode est particulièrement indiqué lorsque l'on dispose d'un échantillon d'effectif faible, ou lorsque l'unité de temps prise en compte est suffisamment petite (le mois) pour que l'on puisse considérer que l'on se situe en temps continu. En revanche, ce qu'il faut éviter, c'est que de nombreux individus connaissent l'événement ou sortent d'observation à un même moment. Ceci était en fait le cas dans notre exemple sur la durée des premiers emplois de plus de trois mois, surtout en début d'observation. Ainsi, nous avons vu que 54 individus ont quitté leur emploi lors du quatrième mois et 33 lors du cinquième (tableau 2.1). On peut donc se demander s'il n'aurait pas mieux valu utiliser les méthodes d'estimation actuarielle, plus adaptées dés lors que les effectifs de personnes soumises au risque sont élevés ou que l'unité de temps est grande.

3. Méthode d'estimation actuarielle

Dans ce point, nous garderons une structure de présentation des méthodes d'estimation actuarielles similaire à celle que nous avons adoptée lors de la présentation des méthodes de KM. Nous nous appuierons, en outre, sur le même exemple concernant la durée des premiers emplois de plus de trois mois des hommes nés entre 1945 et 1964 ayant été interrogés dans le cadre de l'enquête suisse sur la famille.

3.1 Principe d'estimation des différentes distributions

3.1.1 Estimations du risque et du risque cumulé

Les méthodes d'estimation actuarielle reposent sur l'hypothèse selon laquelle le risque instantané $h_i(t)$ est constant tout le long de l'intervalle de temps $[t_i, t_{i+1}[$ (Le Goff, 1994). On considère, en outre, que les échéances, et les sorties d'observation, ont lieu uniformément durant cet intervalle. Ainsi, si l'intervalle correspond à une année, on considérera qu'il y a autant de personnes qui auront connu l'événement en janvier, février, …, ou en décembre. Il en sera de même en ce qui concerne les sorties d'observation. Ceci signifie que les individus qui sortent d'observation ou qui connaissent l'événement durant cet intervalle de temps ont été, **en moyenne**, soumis au risque de connaître l'événement pendant la première moitié de cet intervalle de temps (les six premiers mois de l'année). En conséquence, la population moyenne soumise au risque de connaître l'événement durant l'intervalle de temps correspondra à l'effectif de la population qui n'avait pas encore connu l'événement au début de cet intervalle, diminuée de la moitié des personnes ayant connu l'événement d'une part et de la moitié des personnes étant sortis d'observation d'autre part. Rigoureusement parlant, il ne s'agit pas tout à fait de la population moyenne, mais du nombre de personnes-années présentes, en moyenne, au cours de l'intervalle de temps.

Ainsi, si d_i représente l'effectif des individus connaissant la transition entre t_i et t_{i+1}, c_i le nombre des sorties d'observation au cours de cet intervalle de temps et si N_i est l'effectif des individus soumis au risque en t_i, alors P_i, le nombre de personnes années durant l'intervalle de temps sera :

$$\mathrm{P_i} = N_i - \frac{1}{2}(d_i + c_i) \qquad (3.1)$$

L'estimateur du risque $h(t_i)$ est (Courgeau & Lelièvre, 1989) :

$$h(t_i) = \frac{d_i}{N_i - \frac{1}{2}(d_i + c_i)} \qquad (3.2)$$

Par ailleurs le risque cumulé $H(t_i)$ est estimé à partir des valeurs de $h(t_k)$ où $k(1,2,\ldots,i)$ par :

$$H(t_i) = \sum_{k \le i} \log[1 - h(t_k)] \qquad (3.4)$$

Si les valeurs de $h(t_k)$ sont petites (de l'ordre de 0,01 à 0,05), ce qui est très fréquemment le cas, l'estimation de $H(t_i)$ pourra être simplifiée par :

$$H(t_i) \cong \sum_{k \le i} h(t_k) \qquad (3.5)$$

3.1.2 Estimation de la distribution de la fonction de séjour et de la densité de probabilité

Dans les tables classiques de démographie, par exemple, la table de mortalité, il est d'usage de présenter la série des « quotients ». Ces quotients correspondent à la probabilité de connaître l'événement durant l'intervalle de temps considéré, conditionnellement au fait que les individus n'avaient pas encore connu cet événement au début de l'intervalle de temps. Si q_i représente le quotient de connaître l'événement durant l'intervalle de temps $[t_i, t_{i+1}[$, alors (Pressat, 1983) :

$$q_i = \frac{d_i}{N_i - \frac{1}{2} c_i} \qquad (3.6)$$

et $(1-q_i)$ représentera la proportion de personnes n'ayant pas connu l'événement. L'estimateur non-paramétrique de la fonction de séjour sera estimé par :

$$S(t_i) = S(t_{i-1})(1 - q_i) \qquad (3.7)$$

Par extension, $S(t_i)$ correspondra alors au produit de toutes les probabilités de n'avoir pas connu l'événement entre le début de l'observation et t_i :

$$S(t_i) = \prod_{t_i \le t} (1 - q_i) \qquad (3.8)$$

La variance de $S(t_i)$, qui permet ensuite d'obtenir l'écart type et l'intervalle de confiance, sera estimée de la façon suivante (Blossfeld & Rohwer, 2001):

$$\mathrm{var}\left[S(t_i)\right] = \left[S(t_i)\right]^2 \sum_{k=1}^{i} \frac{\frac{d_k}{N_k}}{\left(1 - \frac{d_k}{N_k}\right) N_{ki}} \qquad (3.9)$$

Par ailleurs, la densité de probabilité $f(t_i)$ sera :

$$f(t_i) = h(t_i) S(t_i) \qquad (3.10)$$

3.2. Estimation actuarielle avec SPSS[18]

3.2.1 Syntaxe SPSS

Nous souhaitons maintenant développer une estimation actuarielle avec SPSS sur les données de durée du premier emploi des hommes nés entre 1945 et 1964. Rappelons que nous disposions précédemment comme variables de base nécessaire à une estimation de Kaplan-Meier, d'une part d'une variable de durée, appelée « *duree* », et d'autre part d'une variable de censure appelée « *censure* ». La première indiquait la durée de l'emploi jusqu'au moment du départ ou jusqu'au moment de l'enquête. La deuxième était égale à 1 si les individus ont quitté leur emploi ou est égale à 0 si le moment de l'enquête intervient avant qu'ils n'aient quitté leur emploi. Ces deux types de variables restent nécessaires dans le cas d'une estimation actuarielle. Néanmoins, dans le cas présent, on peut avoir des durées égales à 0, en ce sens qu'elles indiquent ici que les individus ont connu l'événement ou sont sortis d'observation durant l'intervalle de temps [*0, 1*[. Ceci nous a conduit, en préalable à la mise en œuvre d'une estimation actuarielle, à créer une nouvelle variable de durée, appelée « *duree2* », qui est égale à la variable « *duree* » moins 1.

Les durées sont exprimées en mois écoulés depuis l'embauche. Nous choisissons toutefois de réaliser notre estimation actuarielle en ayant pour unité de temps l'année. En effet, le choix de cette unité de temps permet de limiter le nombre d'informations chiffrées que nous aurons en sortie (cf. tableau 3.1). Une première solution consisterait à créer, préalablement à la mise en œuvre de la procédure SPSS permettant de faire des estimations actuarielles, une nouvelle variable dans laquelle les durées observées seraient converties en année (en tronquant les chiffres après la virgule). Néanmoins, la commande SPSS de l'estimation actuarielle permet directement de choisir l'unité de temps en définissant la largeur de l'intervalle sur laquelle on souhaite obtenir une estimation des différentes grandeurs. Ainsi, nous avons écrit sur un fichier de syntaxe les lignes suivantes :

```
SURVIVAL
  TABLE=duree2
  /INTERVAL=THRU 420 BY 12
  /STATUS=censure(1)
  /PRINT=TABLE
  /PLOTS (SURVIVAL HAZARD)=duree2.
```

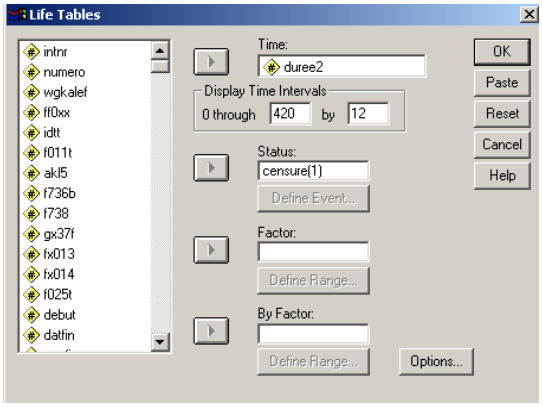
La commande « SURVIVAL » fait appel à une estimation actuarielle. La seconde ligne permet de déclarer la variable de durée « *duree2* ». Dans la troisième ligne, est définie d'une part, la durée maximum pour laquelle on souhaite une analyse et la largeur de l'intervalle de temps que l'on souhaite prendre en compte. Dans le cas présent, nous avons choisi de limiter notre analyse à 420 mois (ce qui représente déjà une durée de 25 années). La largeur de l'intervalle de temps est de 12 mois, c'est-à-dire, une année. STATUS= censure(1) donne l'indice de censure qui est égal à 1 si les individus ont quitté leur emploi, 0 s'ils n'ont pas connu cet événement au moment de l'enquête. Les deux lignes suivantes indiquent que l'on souhaite en sortie la table des résultats ainsi que les graphiques de la fonction de séjour *S(t)* et du risque *h(t)*[19].

[18] Comme dans le cas de la procédure KM, les résultats que nous présentons ici ont été estimés à partir d'un usage de la version 11.5 de SPSS.

[19] Il s'agit bien ici du graphique de la distribution du risque *h(t)* et non du risque cumulé *H(t)*, comme cela était le cas dans la procédure KM.

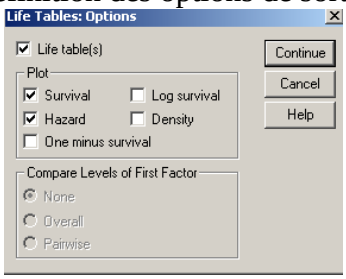
Ces commandes peuvent aussi être appelées par les boîtes de dialogue. Dans le menu *Analyse*, sélectionner *Survival* et *Life Tables*. Choisir ensuite la durée (*Time*), les intervalles de temps désirés (*Display Time Interval*) et indiquer la valeur de l'indice de censure (*status*) pour lequel l'événement se produit (figure 3.1).

Figure 3.1 : Procédure SURVIVAL par la boîte de dialogue : définition des variables de durée et de censure



La touche *options* permet d'indiquer si l'on souhaite une table (Life table en sortie) et de choisir les graphiques que l'on veut obtenir en sortie (figure 3.2). Sont ici demandés la table de survie et les graphiques de *S(t)* et de *H(t)*. D'autres types de graphiques peuvent être sélectionnés.

Figure 3.2 : Procédure SURVIVAL par la boîte de dialogue : définition des options de sortie



3.2.2 Résultats

Chaque ligne de la table des résultats correspond à un intervalle de temps (tableau 3.1). Dans notre exemple, ces intervalles sont de 12 mois. Ainsi, le premier intervalle est 0-11 mois, le deuxième, 12-23 mois, etc. Les colonnes de la table indiquent respectivement :

- 1) *Interval Start Time* : le début de l'intervalle de temps considéré ;

- 2) *Number Entering This Interval* : l'effectif des individus qui n'ont pas encore connu l'événement ou qui ne sont pas encore sortis d'observation avant le début de l'intervalle de temps ;

- 3) *Number Withdrawn during This Interval* : le nombre d'individus qui sortent d'observation (c_i) durant cet intervalle ;

- 4) *Number Exposed to Risk* : l'effectif des individus soumis au risque de connaître l'événement. **Attention**, dans SPSS, il s'agit de l'effectif des personnes présentes au début de l'intervalle de temps diminué seulement de la moitié des personnes étant sorti d'observation au cours de l'intervalle de temps considéré (N_i-$1/2c_i$). Il s'agit donc de l'effectif nécessaire au calcul du quotient $q(t_i)$ et non du risque h(t_i) ;

- 5) *Number of Terminal Events* : le nombre d_i d'individus ayant connu l'événement au cours de l'intervalle de temps ;

- 6) *Proportion terminating* : le quotient $q(t_i)$;

- 7) *Proportion Surviving* : le complémentaire à 1 de $q(t_i)$;

- 8) *Cumulative Proportion Surviving at End* : la fonction de séjour $S(t_{i+1})$;

- 9) *Probability Density* : la densité $f(t_i)$ qui correspond à la probabilité de connaître l'événement entre t_i et $t_i + \Delta t$. Attention, Δt est égal ici à 1 mois. En d'autres termes, $f(t_i)$ sera ici une moyenne mensuelle en divisant la densité de probabilité estimée sur une année par 12 ;

- 10) *Hazard Rate* : le risque $h(t_i)$. Là encore, il s'agit d'un risque moyen mensuel, c'est-à-dire, le risque tel que l'on peut l'estimer selon la formule que nous avons indiqué auparavant, divisé par 12

La suite de la table indique :

- 11) *Interval Start Time* : de nouveau le début de l'intervalle de temps considéré ;

- 12) *Standard Error of the Cumulative Proportion Surviving* : l'écart type de la fonction de séjour $S(t_{i+1})$;

- 13) *Standard Error of the Probability Density* : l'écart type de la densité $f(t_i)$;

- 14) *Standard Error of the Hazard Rate* : l'écart type de la fonction $h(t_i)$.

Dans le cas présent, entre le début de la période d'observation et la fin du 11ième mois, l'effectif des individus présents en début d'intervalle est de 1 100. Aucun individu ne sort d'observation durant cet intervalle alors que 300 hommes quittent leur emploi. La population soumise au risque est de 1100 et le quotient $q(t_i)$ est égal à 300/(1100-0), soit 27,27%. Le complémentaire à 1 de $q(t_i)$ est de 100-27,27 soit de 72,73% et $S(t_{i+1})$ sera égal à 1*0,7273. La densité de probabilité, qui est égale à 0,2727/12, soit 2,27%, correspond à la probabilité d'avoir connu l'événement au cours de l'intervalle de temps. Le risque $h(t_i)$ a pour valeur 300/12 *(1100-300/2), soit 0,0263[20]. Finalement, les valeurs des écarts types de $S(t_{i+1})$, $f(t_i)$ et $h(t_i)$ sont respectivement de 0,134, 0,11 et 0,15. L'intervalle de confiance est alors $h(1)$ = 0,0263 ± 1,96*0,0015 ; le risque est ainsi compris entre 0,0292 et 0,0233.

[20] Du fait que le risque n'est pas une probabilité ou une proportion (il peut être supérieur à 1), on ne peut pas l'exprimer en pourcentage.

L'hypothèse d'un risque constant durant le premier intervalle de temps n'est pas tout à fait vraie, en ce sens que l'on sait par définition que les individus ne quittent leur activité professionnelle qu'à partir du quatrième mois d'activité. En d'autre terme, le risque de départ est nul durant les trois premiers mois. On peut néanmoins négliger cet aspect.

Au début du 12^{ième} mois, il reste 1100-300 = 800 individus soumis au risque de quitter leur emploi. Aucune sortie d'observation n'a lieu entre ce douzième mois et la fin du 23^{ième}, mais 247 individus quittent leur activité professionnelle. Le quotient est égal à 247/800, soit 30,88%. Le complémentaire à 1 de $q(t_i)$ est 100-30,88 soit 69,12%. $S(t_i)$ sera alors 1*0,7273*0,6912 = 0,5027. En d'autres termes, près de 50% des hommes ont quitté leur activité professionnelle dans les deux ans qui ont suivi leur embauche. L'intervalle de confiance de *h(2)* est 0,0304 ± 1,96*0,0019 soit entre 0,0341 et 0,0266.

On remarquera que l'effectif des individus soumis au risque de connaître l'événement[21] n'est pas nécessairement un nombre entier. Par exemple, la table montre que 218,5 individus sont soumis au risque de départ de l'emploi durant l'intervalle de temps qui va du 72^{e} au 83^{e} mois. Durant cet intervalle de temps, un individu est sorti d'observation. Il a donc été soumis au risque de connaître l'événement pendant ½ année et compte ainsi pour ½ personne-année. Dans ce cas, l'effectif des personnes soumises au risque est de 219-1/2 = 218,5 personnes-années.

La table se termine par une estimation de la durée médiane, c'est-à-dire, la durée au cours de laquelle 50% des individus ont quitté leur emploi. Dans le cas présent, la médiane est de 24,24 mois, ce qui signifie que la moitié des personnes a quitté son premier emploi après deux ans d'activité. Comme nous pouvons le constater, ce résultat reste très proche de celui qui avait été obtenu avec la méthode de Kaplan-Meier (cf. tableau 2.3).

Les graphiques de *S(t)* et de *h(t)* sont édités à la suite de la table de survie (Figures 3.2 et 3.3). Toutefois, la représentation de *S(t)* est un peu trompeuse, en ce sens que les concepteurs de la procédure « Survival » de SPSS ont construit le design de ce graphique de manière à ce que *S(t)* soit une fonction en escalier, c'est-à-dire, une fonction discrète[22]. Or, dans le cas présent, il s'agit d'une fonction continue. Quoiqu'il en soit, on peut observer que cette fonction diminue rapidement durant les cinquante premiers mois.

Le graphique de *h(t)* est difficile à lire car les points ne sont pas reliés entre eux. Un graphique Excel reprenant la série des risques estimés (en reliant les points entre eux) permet d'y voir un peu plus clair (figure 3.5). Les risques sont ainsi décroissants entre le 6^{ième} et le 166^{ième} mois. Mais après? Du fait de la faiblesse des effectifs, son évolution paraît aléatoire. Un graphique de *H(t)* est plus intéressant puisqu'il permet de lisser ces effets aléatoires (Figure 3.6). Jusqu'au 50^{ième} mois, le risque de quitter son emploi est constant au cours du temps, puis la pente de forme convexe montre que le risque diminue au cours du temps. Ce résultat va dans le sens de notre hypothèse selon laquelle un premier emploi est une phase transitoire pour un nombre important de jeunes, mais qu'il peut aussi devenir un emploi d'insertion pour un nombre non négligeable d'autres jeunes.

[21] Au sens SPSS c'est-à-dire pour le calcul des quotients, mais cela reste vrai dans la définition de l'effectif des personnes années soumises au risque de connaître l'événement.
[22] C'est plutôt dans la procédure KM qu'une fonction en escalier serait parfaitement adéquate.

Tableau 3.1 : Listing analyse des emplois

This subfile contains: 1100 observations

Life Table

Survival Variable DUREE2

Intrvl Start Time	Number Entrng this Intrvl	Number Wdrawn During Intrvl	Number Exposd to Risk	Number of Termnl Events	Propn Termi-nating	Propn Sur-viving	Cumul Propn Surv at End	Proba-bility Densty	Hazard Rate
.0	1100.0	.0	1100.0	300.0	.2727	.7273	.7273	.0227	.0263
12.0	800.0	.0	800.0	247.0	.3088	.6912	.5027	.0187	.0304
24.0	553.0	.0	553.0	147.0	.2658	.7342	.3691	.0111	.0255
36.0	406.0	2.0	405.0	102.0	.2519	.7481	.2761	.0077	.0240
48.0	302.0	2.0	301.0	57.0	.1894	.8106	.2238	.0044	.0174
60.0	243.0	.0	243.0	24.0	.0988	.9012	.2017	.0018	.0087
72.0	219.0	1.0	218.5	24.0	.1098	.8902	.1796	.0018	.0097
84.0	194.0	2.0	193.0	23.0	.1192	.8808	.1582	.0018	.0106
96.0	169.0	1.0	168.5	13.0	.0772	.9228	.1460	.0010	.0067
108.0	155.0	2.0	154.0	14.0	.0909	.9091	.1327	.0011	.0079
120.0	139.0	3.0	137.5	10.0	.0727	.9273	.1231	.0008	.0063
132.0	126.0	6.0	123.0	7.0	.0569	.9431	.1160	.0006	.0049
144.0	113.0	6.0	110.0	6.0	.0545	.9455	.1097	.0005	.0047
156.0	101.0	6.0	98.0	6.0	.0612	.9388	.1030	.0006	.0053
168.0	89.0	9.0	84.5	.0	.0000	1.0000	.1030	.0000	.0000
180.0	80.0	6.0	77.0	1.0	.0130	.9870	.1017	.0001	.0011
192.0	73.0	4.0	71.0	3.0	.0423	.9577	.0974	.0004	.0036
204.0	66.0	6.0	63.0	3.0	.0476	.9524	.0927	.0004	.0041
216.0	57.0	7.0	53.5	1.0	.0187	.9813	.0910	.0001	.0016
228.0	49.0	4.0	47.0	.0	.0000	1.0000	.0910	.0000	.0000
240.0	45.0	2.0	44.0	1.0	.0227	.9773	.0889	.0002	.0019
252.0	42.0	4.0	40.0	1.0	.0250	.9750	.0867	.0002	.0021
264.0	37.0	9.0	32.5	1.0	.0308	.9692	.0840	.0002	.0026
276.0	27.0	3.0	25.5	.0	.0000	1.0000	.0840	.0000	.0000
288.0	24.0	2.0	23.0	.0	.0000	1.0000	.0840	.0000	.0000
300.0	22.0	5.0	19.5	.0	.0000	1.0000	.0840	.0000	.0000
312.0	17.0	3.0	15.5	.0	.0000	1.0000	.0840	.0000	.0000
324.0	14.0	6.0	11.0	.0	.0000	1.0000	.0840	.0000	.0000
336.0	8.0	3.0	6.5	.0	.0000	1.0000	.0840	.0000	.0000
348.0	5.0	2.0	4.0	.0	.0000	1.0000	.0840	.0000	.0000
360.0	3.0	1.0	2.5	.0	.0000	1.0000	.0840	.0000	.0000
372.0	2.0	.0	2.0	.0	.0000	1.0000	.0840	.0000	.0000
384.0	2.0	1.0	1.5	.0	.0000	1.0000	.0840	.0000	.0000
396.0	1.0	1.0	.5	.0	.0000	1.0000	.0840	.0000	.0000

The median survival time for these data is 24.24

Intrvl Start Time	SE of Cumul Sur-viving	SE of Proba-bility Densty	SE of Hazard Rate
.0	.0134	.0011	.0015
12.0	.0151	.0010	.0019
24.0	.0145	.0009	.0021
36.0	.0135	.0007	.0024
48.0	.0126	.0006	.0023
60.0	.0121	.0004	.0018
72.0	.0116	.0004	.0020
84.0	.0110	.0004	.0022
96.0	.0107	.0003	.0019
108.0	.0103	.0003	.0021
120.0	.0100	.0003	.0020
132.0	.0098	.0002	.0018
144.0	.0096	.0002	.0019
156.0	.0094	.0002	.0021
168.0	.0094	.0000	.0000
180.0	.0093	.0001	.0011
192.0	.0093	.0002	.0021
204.0	.0092	.0002	.0023
216.0	.0092	.0001	.0016
228.0	.0092	.0000	.0000
240.0	.0092	.0002	.0019
252.0	.0093	.0002	.0021
264.0	.0093	.0002	.0026
276.0	.0093	.0000	.0000
288.0	.0093	.0000	.0000
300.0	.0093	.0000	.0000
312.0	.0093	.0000	.0000
324.0	.0093	.0000	.0000
336.0	.0093	.0000	.0000
348.0	.0093	.0000	.0000
360.0	.0093	.0000	.0000
372.0	.0093	.0000	.0000
384.0	.0093	.0000	.0000
396.0	.0093	.0000	.0000

Figure 3.3 : Fonction de séjour en emploi (estimation actuarielle)

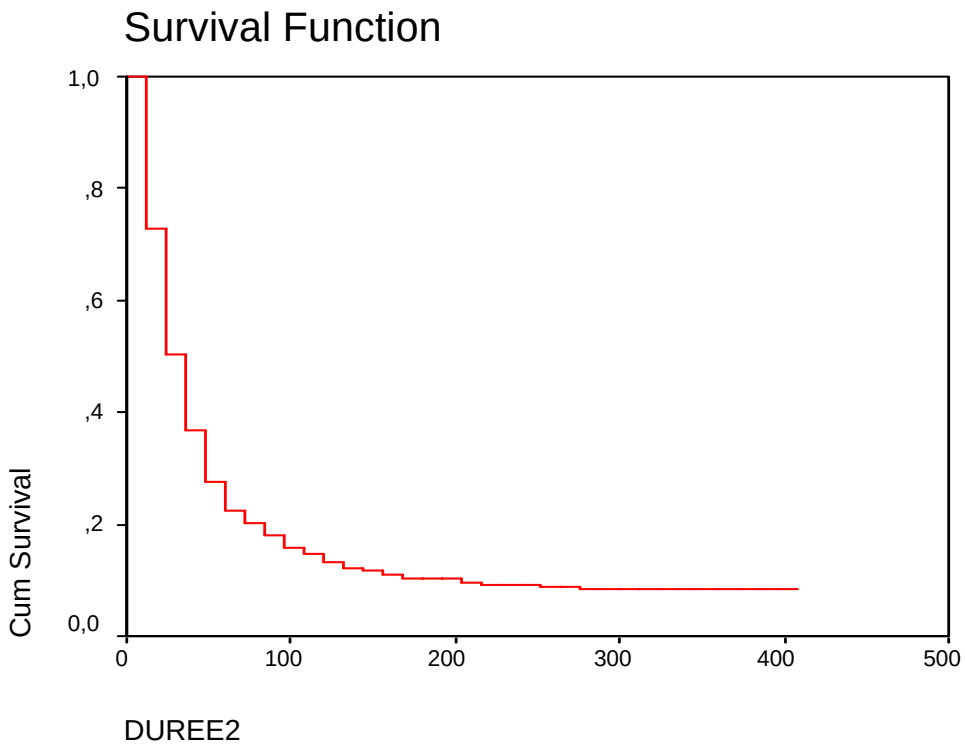


Figure 3.4 : Fonction de risque de départ en emploi
(estimation actuarielle - graphique SPSS)

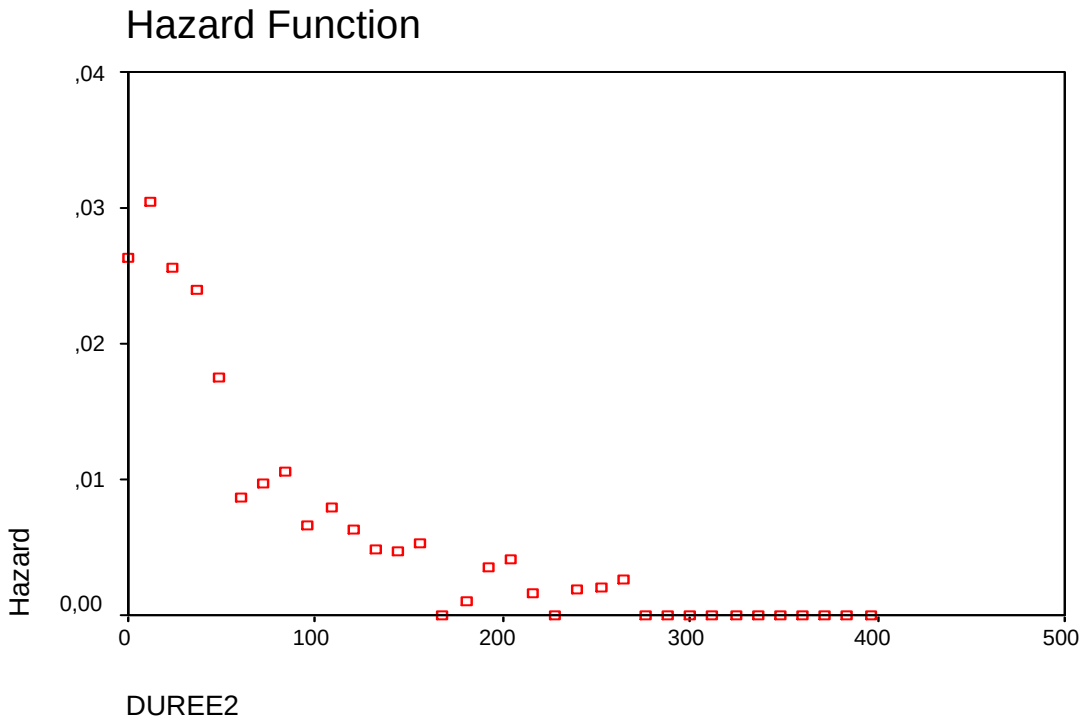


*Figure 3.5 : F*onction de risque de départ en emploi (reconstituée avec Excel)

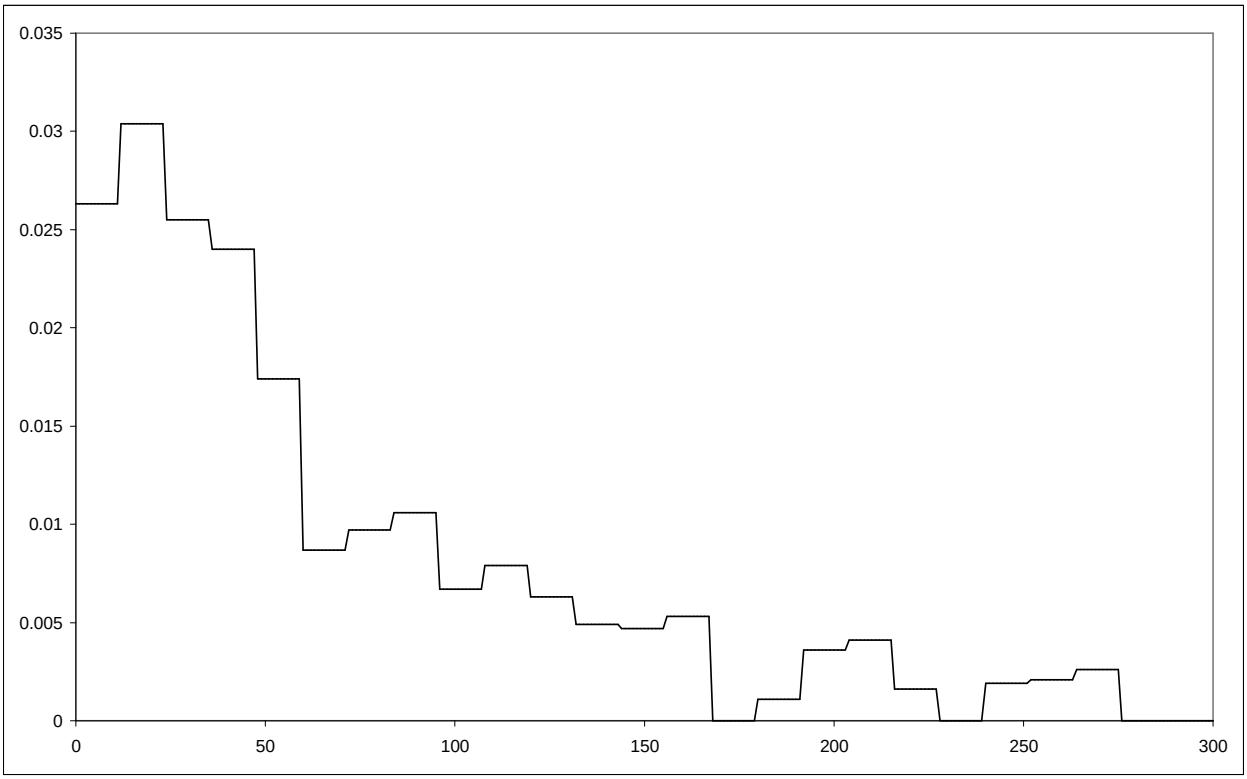
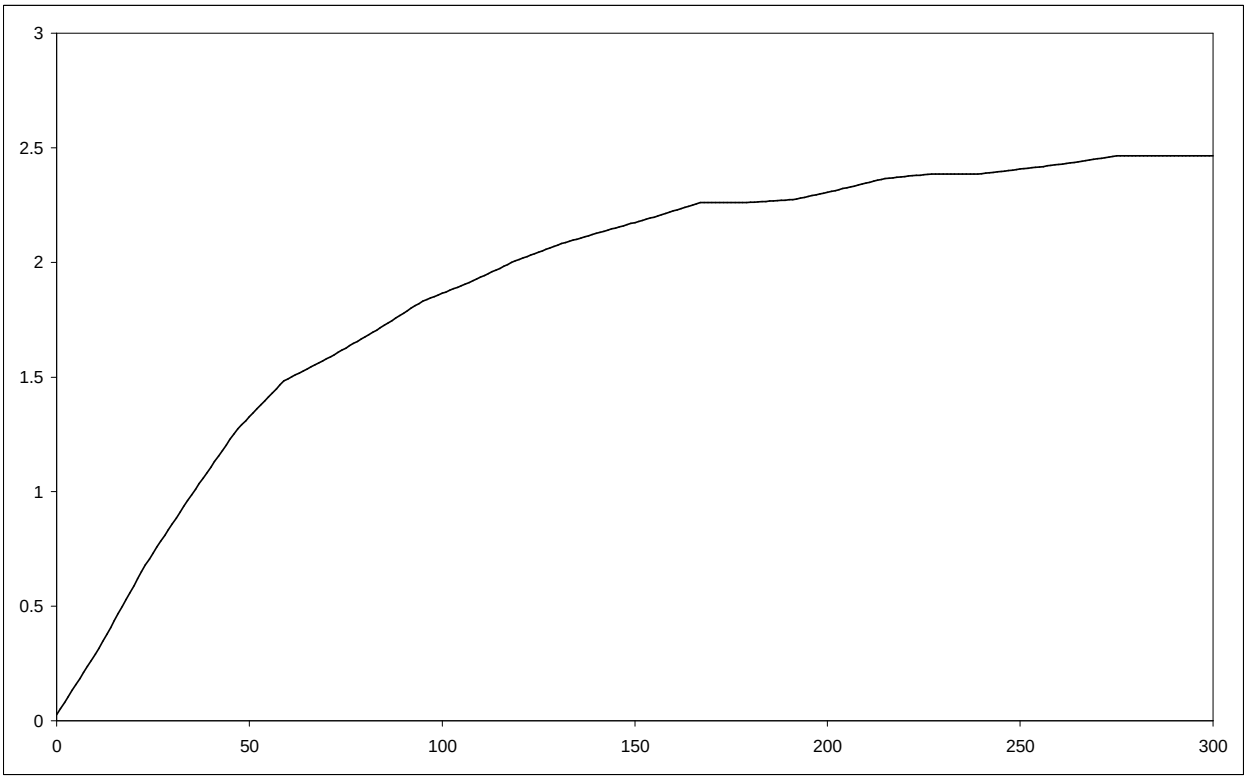


Figure 3.6 : Risques cumulés de départ d'emploi



3.3. Comparaison de plusieurs sous populations

3.3.1 Deux sous-populations

Contrairement à la procédure KM dans laquelle trois tests étaient disponibles, la procédure « *survival* » de SPSS n'offre qu'un seul test. Il s'agit du test de Wilcoxon (Gehan), qui est un test analogue aux tests précédents. Il est plutôt sensible à des différences en début de période d'observation.

Comme cela était le cas dans l'estimation de Kaplan-Meier, notre intérêt porte sur la comparaison des différentes distributions concernant la durée du premier emploi de plus de trois mois entre les hommes nés entre 1945 et 1954 et ceux nés entre 1955 et 1964. Nous sommes donc ici en présence d'un facteur (*factor*) de distinction qui est la cohorte de naissance. Cette variable, que nous avons appelée « coho » est égale à 1 si la cohorte de naissance est 1945-54 et à 2 si la cohorte de naissance est 1955-64. Cette comparaison a été réalisée au travers de la syntaxe suivante :

```
SURVIVAL
  TABLE=duree2 BY coho(1 2)
  /INTERVAL=THRU 420 BY 12
  /STATUS=censure(1)
  /PRINT=TABLE
  /PLOTS ( SURVIVAL HAZARD )=duree2 BY coho
  /COMPARE=duree2 BY coho.
```

Dans la seconde ligne est déclarée la variable de « durée2 ». La mention « BY coho » indique que ce sont les distributions des durées d'emploi pour chacune des deux cohortes décennales qui seront données en sortie. Cependant, on doit préciser à quelles sous populations on va s'intéresser. Ici, comme l'on prend en compte les deux cohortes, on écrira « *coho(1 2)* » où les deux nombres à l'intérieur de la parenthèse, ici *(1 2)*, désignent respectivement les rangs minimal et maximal de la variable prise en tant que facteur de distinction. Les spécifications de la largeur de l'intervalle et de l'indice de censure restent inchangées, de même que les commandes PRINT et PLOT. La dernière ligne permet de demander le test de comparaison entre les deux cohortes.

Ces commandes peuvent être exécutées directement à partir de la boîte de dialogue. Ainsi, sélectionner *Analyse*, puis *Survival* et *Life Tables*. Spécifier ensuite la variable de durée (*Time*), ici dure2, les intervalles de temps désirés (*Display Time Interval*), ici 12 et indiquer la valeur de l'indice de censure (*status*) pour lequel l'événement se produit (*Define Events*). Sélectionner également la variable de premier facteur (*coho*) et définir les rangs (*min; max*), ici 1 et 2. Afin d'obtenir les graphiques, permettant la comparaison des sous-populations ou pour supprimer des tables de survie, aller sur *Options*. Il est également possible de comparer différents niveaux du premier facteur (*Compare Levels of First Factor*).

b) Résultats

Le résultat du test et les graphiques sont édités à la suite des tables dans un fichier de sortie. Dans le cas présent, comme l'on compare deux sous populations, le test de Wilcoxon doit être comparé à un χ^2 à 1 degré de liberté. Le test est ici significatif au seuil de 5%, ce qui signifie que la distribution des événements diffère entre les cohortes (tableau 3.2). Ce tableau se termine en récapitulant le nombre d'individus sortis d'observation ou qui connaissent l'événement (ici, quitter son premier emploi) ainsi que le score moyen (*mean score*). Ce score

moyen est calculé en comparant les individus les uns par rapport aux autres, c'est-à-dire, en incrémentant le score de 1 pour chaque individu ayant un temps de survie plus long qu'un autre, et en décrémentant de 1 pour chaque individu ayant un temps de survie plus court qu'un autre.

Les figures 3.7 et 3.8 représentent respectivement les graphiques SPSS de la fonction de séjour *S(t)* et de la fonction de risque *h(t)* pour chacune des deux cohortes. Les comportements des deux cohortes apparaissent relativement semblables jusqu'au 50^{ième} mois, puis l'écart augmente légèrement. La durée du premier emploi pour la cohorte 1955-1964 est plus courte que pour la cohorte 1945-1954.

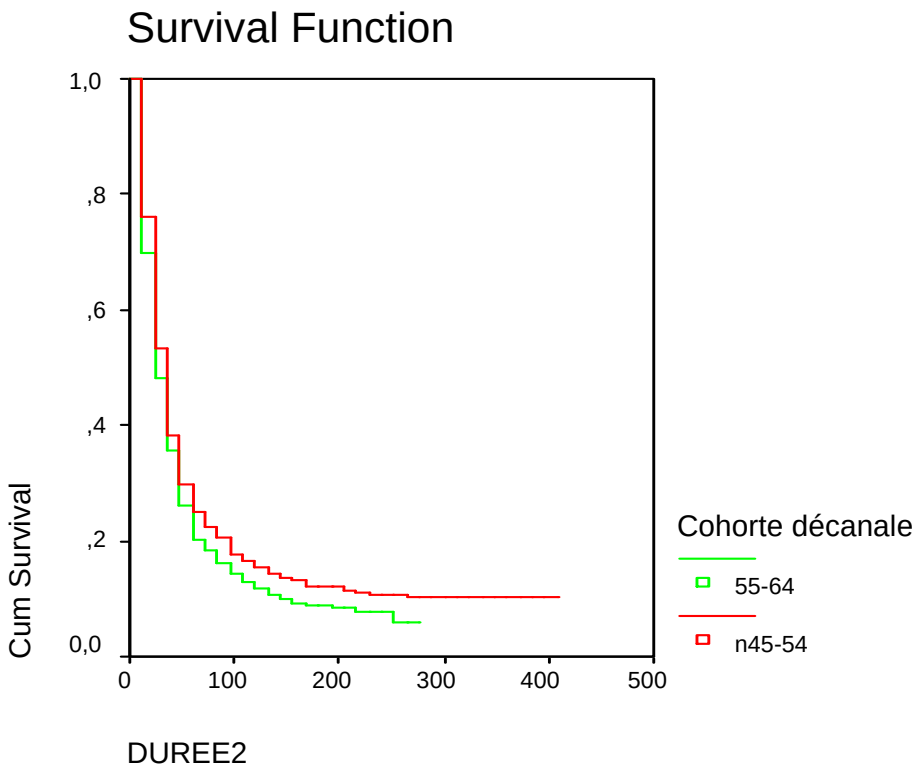
Tableau 3.2 : Résultat du test de Wilcoxon (Gehan) de comparaison des distributions de séjour

Survival Variable DUREE2
grouped by COHO Cohorte décanale

Overall comparison statistic 4,171 D.F. 1 Prob. ,0411

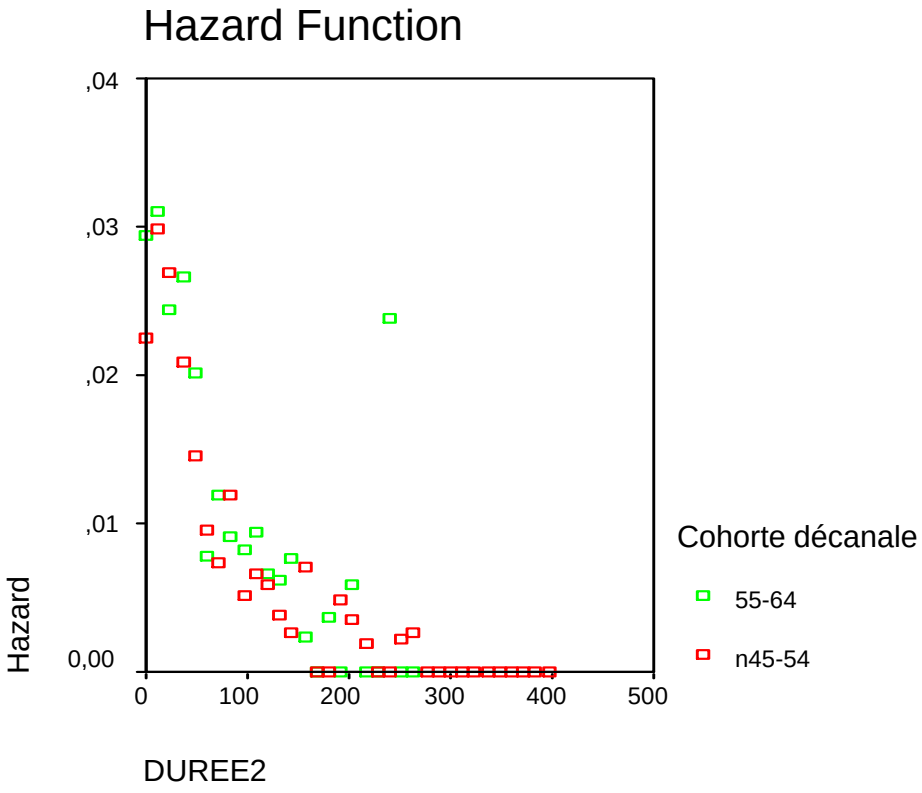
Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
1	n45-54	480	429	51	10,63	44,4042
2	55-64	620	562	58	9,35	-34,3774

Figure 3.7 : Fonction de séjour en emploi selon la cohorte de naissance (estimation actuarielle)



;

Figure 3.8 : Risque de départ d'emploi selon la cohorte de naissance (estimation actuarielle)



3.3.2 Plus de deux sous populations

Dans ce point, nous allons décrire la procédure que nous avons utilisée pour analyser les différences entre les cohortes quinquennales sur la base d'estimations actuarielles. Rappelons que les modalités 6, 5, 4, et 3 de la variable *akl5* indiquent que les individus appartiennent respectivement aux cohortes 1945-49, 1950-54, 1955-59 et 1960-64. L'ensemble des commandes suivantes a été écrit dans une feuille de syntaxe :

```
SURVIVAL
  TABLE=duree2 BY akl5 (3 6)
  /INTERVAL=THRU 420 BY 12
  /STATUS=censure(1)
  /PRINT=TABLE
  /PLOTS ( SURVIVAL HAZARD )=duree2 BY akl5
  /COMPARE=duree2 BY akl5.
```

Par rapport à la syntaxe telle que celle-ci avait été écrite dans le cas de la comparaison des deux cohortes décanales, à chaque fois la variable « *coho* » a été remplacée par la variable « *akl5* », avec la mention (3 6) dans la deuxième ligne pour signifier que l'on s'intéresse à l'ensemble des modalités comprises entre 3 et 6. Dans le cas présent, il n'y a qu'un test de comparaison entre les sous populations et la population totale. La statistique obtenue est comparée avec un χ^2 à autant de degrés de liberté qu'il y a de sous populations, moins 1. Ici, le nombre de degrés de liberté sera égal à 3 (tableau 3.3). Dans le cas présent, le test n'est pas significatif. En d'autres termes, il n'y a pas de différences dans les rythmes de départ d'emploi entre les différentes cohortes quinquennales.

Tableau 3.3 : Résultat du test dans le cas de plus de deux sous-populations
Comparaison de l'ensemble des sous populations

```
Comparison of survival experience using the Wilcoxon (Gehan) statistic
   Survival Variable  DUREE2
          grouped by  AKL5       5-Jahres-Altersklasse in Bezug auf AGE

  Overall comparison    statistic        4,614  D.F.     3   Prob.   ,2023

  Group  label                Total N   Uncen     Cen  Pct Cen  Mean Score

      3  30 - 34 ans              287     256      31    10,80    -22,0836
      4  35 - 39 ans              333     306      27     8,11    -44,9730
      5  40 - 44 ans              252     226      26    10,32     57,9722
      6  45 - 49 ans              228     203      25    10,96     29,4079
```

L'ajout d'une nouvelle ligne à la syntaxe précédente :

```
../CALCULATE PAIRWISE.
```

permet d'obtenir une série de tests dans lesquels les sous populations sont comparées deux à deux. Dans le cas présent, comme il y a quatre sous-populations, il y a six séries de tests (cf. tableaux 3.4). Comme dans le cas de l'estimation de KM, aucun test n'est ici significatif, vraisemblablement en raison des faibles effectifs de chacune des cohortes.

Tableau 3.4 : Résultat du test dans le cas de plus de deux sous-populations
Comparaison de chaque sous population deux à deux

Comparison of survival experience using the Wilcoxon (Gehan) statistic
Survival Variable DUREE2
grouped by AKL5 5-Jahres-Altersklasse in Bezug auf AGE

Pairwise comparison statistic ,214 D.F. 1 Prob. ,6437

Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
3	30 - 34 ans	287	256	31	10,80	7,1568
4	35 - 39 ans	333	306	27	8,11	-6,1682

Pairwise comparison statistic 2,250 D.F. 1 Prob. ,1336

Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
3	30 - 34 ans	287	256	31	10,80	-18,8223
5	40 - 44 ans	252	226	26	10,32	21,4365

Pairwise comparison statistic 3,771 D.F. 1 Prob. ,0521

Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
4	35 - 39 ans	333	306	27	8,11	-23,5886
5	40 - 44 ans	252	226	26	10,32	31,1706

Pairwise comparison statistic ,797 D.F. 1 Prob. ,3719

Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
3	30 - 34 ans	287	256	31	10,80	-10,4181
6	45 - 49 ans	228	203	25	10,96	13,1140

Pairwise comparison statistic 1,808 D.F. 1 Prob. ,1787

Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
4	35 - 39 ans	333	306	27	8,11	-15,2162
6	45 - 49 ans	228	203	25	10,96	22,2237

Pairwise comparison statistic ,199 D.F. 1 Prob. ,6557

Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
5	40 - 44 ans	252	226	26	10,32	5,3651
6	45 - 49 ans	228	203	25	10,96	-5,9298

3.4. Précisions sur la commande « SURVIVAL » de SPSS

Les instructions qui suivent permettent de préciser les informations souhaitées dans le fichier de sortie :

/PRINT =TABLE, NOTABLE

PRINT TABLE est utilisé pour obtenir la table de survie de l'estimation actuarielle (une pour chaque sous population). PRINT NOTABLE supprime toute table en sortie. En l'absence de l'instruction PRINT, seule est donnée la table de survie.

/PLOT = ALL, SURVIVAL, HAZARD, OMS, DENSITY, LOGSURV

Si l'option PLOT n'est pas spécifiée, aucun graphique n'est tracé en sortie. Cinq possibilités de graphique peuvent être spécifiées à partir de cette commande. Les options HAZARD et SURVIVAL donnent respectivement les graphiques du risque *h(t)* de connaître l'événement au cours du temps et de la fonction de séjour *S(t)*, alors que l'option OMS (*One Minus Survival*) permet d'obtenir un graphique de la proportion cumulée d'individus ayant connu l'événement. DENSITY correspond à un graphique de la densité de probabilité *f(t)* et l'option LOGSURV donne un graphique de *S(t)* sur une échelle logarithmique. PLOT ALL permet d'avoir en sortie tous les graphiques, de même que la seule spécification de PLOT sans autre précision.

/CALCULATE=[EXACT,CONDITIONAL,APPROXIMATE][PAIRWISE,COMPARE]

CALCULATE contrôle les comparaisons des fonctions de survie pour les sous-populations spécifiées par la commande COMPARE (SPSS, 1993). Seuls les mots-clés EXACT, APPROXIMATE et CONDITIONAL peuvent être spécifiés. Les mots-clés PAIRWISE et COMPARE peuvent être utilisés avec EXACT, APPROXIMATE et CONDITIONAL.

Le mot-clé EXACT est utilisé par défaut. Cette méthode est en fait possible si toutes les données sont en mémoire simultanément. Ainsi, les comparaisons « exact » ne sont pas praticables avec de grands échantillons. Les comparaisons avec le mot-clé APPROXIMATE sont appropriées pour des données agrégées. Cette méthode est basée sur l'idée que tous les événements ont lieu au milieu de l'intervalle de temps. CONDITIONNAL effectue des comparaisons s'il n'y a pas suffisamment de mémoire disponible pour les comparaisons EXACT. La commande PAIRWISE est utile pour effectuer des comparaisons deux à deux et COMPARE permet de faire des comparaisons entre les sous-populations. Il est à noter que la commande WRITE (cf. plus loin) ne peut pas être utilisée avec COMPARE.

/MISSING=GROUPWISE, LISTWISE, INCLUDE

L'instruction MISSING contrôle le traitement des données manquantes (SPSS for Windows, 1993). Les valeurs négatives sont automatiquement traitées comme des données manquantes. GROUPWISE est utilisé par défaut. Avec l'option GROUPWISE, les données manquantes sont exclues de tout calcul impliquant la variable étudiée (*Exclude missing value groupwise*). Avec l'option LISTWISE, les données manquantes de n'importe quel type de variables sont exclues de l'analyse (*Exclude missing value listwise*). La commande INCLUDE permet d'inclure les données manquantes dans l'analyse. GROUPWISE et LISTWISE sont mutuellement exclusifs, mais chaque commande peut être utilisée avec INCLUDE.

/WRITE = NONE, TABLES, BOTH

Cette instruction écrit des données des tables de survie en un fichier de données « .SAV », ce qui peut être utile, par exemple, si l'on souhaite faire des graphiques. Il en est de même si l'on utilise la commande NONE. Avec la commande TABLES, toutes les données de la table de survie sont écrites sur un nouveau fichier. BOTH permet d'obtenir à la fois les données de la table de survie, mais également le nom des variables, l'étiquette des variables et les valeurs des étiquettes. Quand WRITE est utilisé, une procédure de sortie (PROCEDURE OUTPUT OUTFILE) doit précéder la commande SURVIVAL. Cette procédure de sortie permet de spécifier le fichier dans lequel vont être écrites les données.

BY

La mention « TABLE = duree2 BY coho (1 2) indique ici que l'on souhaite des tables qui concernent la distribution de la durée de l'emploi selon les cohortes (variable coho). Cette variable est considérée en tant que premier facteur. L'instruction « BY » peut être utilisée pour déclarer un second facteur. Par exemple, si l'on avait les emplois des femmes et que l'on souhaiterait distinguer les hommes des femmes, on pourrait écrire : « TABLE = duree2 BY coho (1 2) BY sexe (1 2) ». Le même procédé peut être réalisé avec l'option « /PLOT » et l'option « /COMPARE ».

*** * ***

Dans ce point, nous avons vu comment utiliser les méthodes d'estimation actuarielle avec SPSS. Rappelons ce que nous avions souligné dans l'introduction de cet article, à savoir que l'usage de ce type de méthode est particulièrement indiqué lorsque l'on dispose de larges effectifs, où les individus connaissent l'événement à chaque instant ou bien si les durées ont été mesurée sur une unité de temps large. Dans notre exemple sur la durée des premiers emplois de plus de trois mois, cela était le cas, en début d'observation. De ce point de vue, la méthode actuarielle semble plus « adaptée » à notre exemple sur la durée des premiers emplois que la méthode d'estimation de Kaplan-Meier.

4. Conclusion

Dans cet article, notre intérêt a porté sur l'application des méthodes non-paramétriques de l'analyse des biographies à partir des possibilités disponibles dans SPSS. Ce package permet ainsi de réaliser des estimations de Kaplan-Meier et des estimations actuarielles. En sortie, nous obtenons des tables ou des graphiques qui permettent de mieux comprendre la distribution d'un événement du parcours de vie au cours du temps. En outre, pour chacune des deux méthodes, il y a la possibilité de comparer la distribution de ces échéances entre deux ou plusieurs sous-populations.

En termes de stratégie de modélisation, les méthodes non-paramétriques doivent être considérées comme un préalable à d'autres analyses de type semi-paramétriques (modèle de Cox) ou paramétriques (modèle exponentiel, de Weibull, loglogistique), ainsi que les méthodes en temps discret. En effet, aussi bien les méthodes de Kaplan-Meier que les méthodes actuarielles permettent de voir l'évolution de la distribution du risque au cours du temps. Ceci peut aider à décider du choix de cette distribution, lorsque l'on veut mettre en œuvre une estimation paramétrique. Le constat d'un risque toujours croissant ou décroissant

indique qu'une distribution de Weibull ou une distribution de Gompertz peut être ajustée aux données, alors qu'un risque croissant puis décroissant laisse entendre qu'une distribution log-logistique ou log-normale serait plus adéquate. En outre, la comparaison des distributions entre différentes sous-populations permet de vérifier les hypothèses qui sont faites concernant les différences de risque en fonction des caractéristiques des individus. Ainsi, aussi bien l'hypothèse de risque proportionnel qui est faite dans les modèles semi-paramétriques et dans un grand nombre de modèles paramétriques que l'hypothèse de temps de sortie accélérée peuvent être testés à partir de méthodes graphiques basées sur des estimations non-paramétriques (Courgeau & Lelièvre 1989, Blossfeld & Rohwer 2001 ; Wu, 1990).

Ajoutons pour terminer que les méthodes de Kaplan-Meier ou actuarielles peuvent être utilisées dans des cas où il y a plusieurs échéances possibles, par exemple, lorsque l'on distingue les causes de mortalité, ou les différentes raisons de départ d'emploi (Le Goff, 1997). Une estimation non-paramétrique pour chacune des causes est simple en ce sens qu'il suffit de procéder à plusieurs estimations les unes à la suite des autres, en changeant seulement l'indice de censure. Dans le cas de risques concurrents, il est important de souligner que l'hypothèse sous-jacente à l'utilisation de l'une ou de l'autre de ces deux méthodes non-paramétriques est que la distribution de la fonction de séjour ou du risque pour une cause est estimée en l'absence de toute autre cause (Courgeau & Lelièvre, 1989).

Références bibliographiques

Allignol A., Beyersmann J. & Schumacher M. (2008). mvna: An R package for the Nelson-Aalen estimator in multistate models. *R News*, 8(2):48-50.

Allison Paul (1995). *Survival Analysis Using the SAS® System*. Cary: SAS Campus Drive.

Blossfeld Hans-Peter & Rohwer Goetz (2001). *Techniques of Event History Modelling. New Approaches to Causal Analysis*. 2nd ed. Mahwah, New-Jersey : Lawrence Erlbaum Associates [1er ed. 1995].

Bocquier Philippe (1996). *L'analyse des enquêtes biographiques à l'aide du logiciel Stata*. Paris : CEPED.

Boltanski Luc & Chiappello Eve (1999). *Le nouvel esprit du capitalisme*. Paris : Gallimard, nrf essais.

Courgeau Daniel & Lelièvre Eva (1989). *Analyse démographique des biographies*. Paris : INED.

Collett Dave (1994). *Modelling Survival Data in Medical Research*. London: Chapman and Hall.

Gabadinho Alexis (1998), *L'enquête suisse sur la famille*, Berne, OFS.

Gabadinho Alexis & Wanner Philippe (1999), *Fertility and Family Surveys in Countries of the ECE Region. Standard Country Report. Switzerland*, Geneva, United Nations Economic Commission for Europe, United Nations Population Fund, Economics Studies n°10m.

Kleinbaum David G. (1996). *Survival Analysis. A Self-Learning Text*. New-York, Berlin : Springer Verlag.

Le Goff Jean-Marie (1994). *Pratique de l'analyse démographique des biographies. Utilisation des procédures Lifetest, Lifereg et Phreg de SAS*. Caen : Institut du Longitudinal, CNRS. Document d'appui pour l'atelier informatique de l'école d'été *Approches et méthodes longitudinales en sciences sociales.*

Le Goff Jean-Marie (1997). « Mobilité des jeunes à l'issue de leur premier emploi stable ». *Population*. 3 : 545-570.

Lelièvre Eva & Bringé Arnault (1998). *Manuel pratique pour l'analyse statistique des biographies : présentation des modèles de durée et utilisation des logiciels SAS, TDA, STATA*. Paris : INED.

Macura Miroslav, Beets Gijs & Burkimsher Marion (2002), « Fertility and Partnership : Why the FFS and what did we learn for it? », in Macura Miroslaw and Beets Gijs : *Dynamics fertility and partnership in Europe. Insight and lessons from corporative research*. Geneva : United Nations Economic Commission for Europe.

Mills Melinda (1999). *Construction of Input Data for Log-Linear Models of Event Histories.* Groningen : Population Research Centre-University of Groningen. Working Paper $n^0$3.

Pressat Roland (1983), *L'analyse démographique*. Paris : PUF (3^e ed.)

SPSS (1993). *SPSS for Windows : Advanced Statistics, Release 6.0.* Chicago. SPSS Inc.

Therneau Terry. M & Grambsch Patricia M. (2000). *Modelling Survival Data : Extending the Cox Model*. Berlin-New York : Springer Verlag.

Vaupel Jim, Manton Kenneth.G. & Stallard Eric. (1979). « The impact of heterogeneity in individual frailty on the dynamics of mortality », *Demography* 16: 439-454.

Vaupel Jim & Yashin Anatoli (1985). « Heterogeneity's Ruses : Some Surprising Effects of Selection on Population Dynamics». *The American Statistician*, 39 (3) : 176-185.

Wu Lawrence.(1990). Simple Graphical goodness-of-Fit Tests for Hazard Rate. In Mayer Karl Ulrich & Tuma Nancy (Eds). *Event History Analysis in Life Course Research*. Madison: The University of Wisconsin Press: 184-199.