

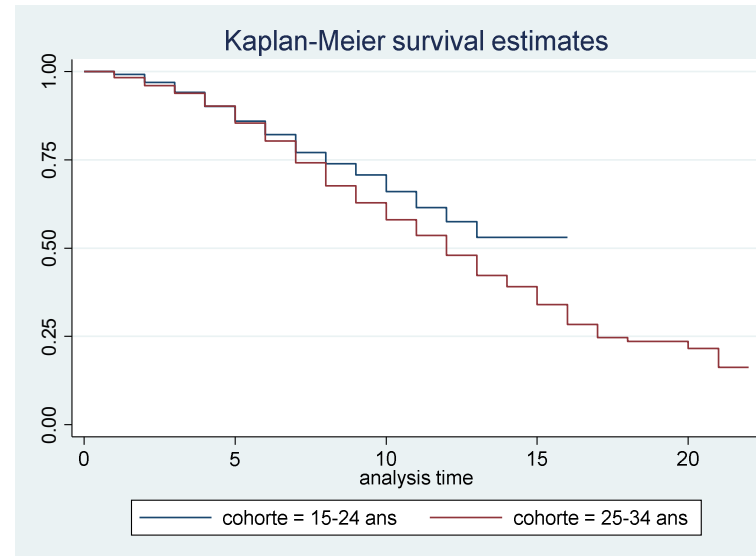
Initiation à l'Analyse de survie



KABORE T.A. Geraude
Démographe
Mail: kaboralimata@gmail.com
Tel: (+226) 73382412

OUEDRAOGO Houdou
Statisticien Economiste
Mail: houdbarack@gmail.com
Tel: (+226) 70129773

COULIBALI Moussa
Démographe
Mail:
coulibalymoussa330@gmail.com
Tel: (+226) 72555573



Année académique_2021-2022

Objectifs

- ▶ Connaître les principes de base de l'analyse de survie
- ▶ Connaître les conditions d'utilisation et interpréter les résultats de la regression de Cox



Plan de présentation

Introduction à l'analyse de survie

Notions d'événement et de troncatures

Données biographiques

Méthode d'analyse descriptive: estimateurs de Kaplan-Meier

Méthode d'analyse explicative : régression semi-paramétrique de Cox

Résumé des notions essentielles

Généralités

- ▶ Jusque là, nous connaissions analyser des données transversales (information obtenues à un moment donnée), sur une période de temps assez courte, souvent l'année civile. La particularité de cette branche de la statistique est que la variable Y à analyser correspond à la durée d'un processus ou à une durée avant la survenue d'un événement (durée de mariage, durée de survie de patients ayant eu un cancer du cerveau).
- ▶ Dorénavant, sachons qu'il existe des données longitudinales qui sont des données relatives à l'évolution d'un phénomène dans le temps au niveau individuel, observée périodiquement pendant un laps de temps donné, permettant l'analyse des biographies.
- ▶ Importance de tenir compte du facteur temps dans l'analyse (*C'est la particularité de l'analyse des biographies*)

Généralités

- ▶ Il s'agit aussi d'expliquer l'occurrence (apparition dans le temps ou dans l'espace) d'un événement au cours du temps (accès à l'emploi, accès à la propriété, accès au logement, sortie d'une cohorte, entrée dans les rapports sexuelles, etc.)
- ▶ Analyser l'interaction entre les événements (fumer la cigarette et avoir le cancer, le mariage et la naissance du premier enfant, etc.)
- ▶ Faire une analyse causale (ex. le contexte social influence le comportement des femmes en matière de fécondité, les divorces cause des abandons scolaires des enfants, etc.)
- ▶ Il faut se familiariser avec les dénominations suivantes: l'analyse des transitions, des durées, de survie, causale.

Notion de cause à effet

- ▶ L'événement A est la cause l'événement B si :
- ▶ A et B sont reliés de manière fonctionnelle
- ▶ A précède B dans le temps
- ▶ La variable explicative et la variable à expliquer doivent être situées dans le temps
- ▶ La variable explicative précède le phénomène à étudier, n'oubliez pas on est dans le temps
- ▶ Le principe de la priorité temporelle de la cause sur l'effet doit être respecté

Rôles de l'analyse des biographies

L'analyse des biographies permet de traiter des données longitudinales censurées. Ces données peuvent notamment être obtenues dans des enquêtes comportant un volet rétrospectif, décrivant l'histoire résidentielle, familiale ou professionnelle de l'individu.

On peut alors étudier l'éventuelle survenue d'un évènement dans le cycle de vie d'un individu en fonction de ses caractéristiques.

Exemples

- ▶ Contrôle des constantes des malades toutes les deux heures



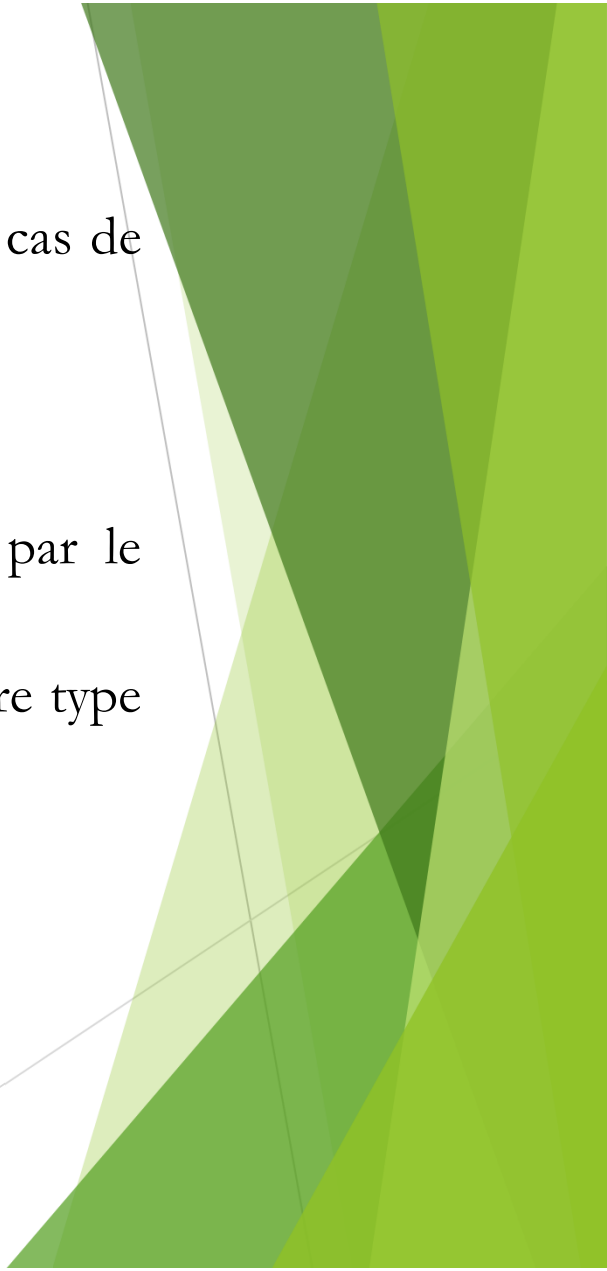
Notions d'événement et de troncatures

Un événement est la transition d'un état qualitatif à un autre. C'est le cas de l'enfance qui est la transition de l'état d'enfant à celui d'adulte

NB: La transition peut se faire vers un seul ou plusieurs états.

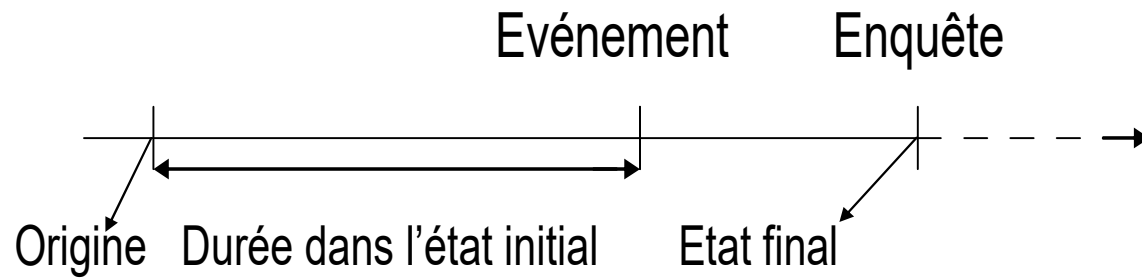
Exemple: La rupture du mariage peut se faire par le divorce ou par le veuvage

ou encore l'instruction peut se faire par la scolarisation ou par un autre type d'éducation



Notions d'événement et de troncatures

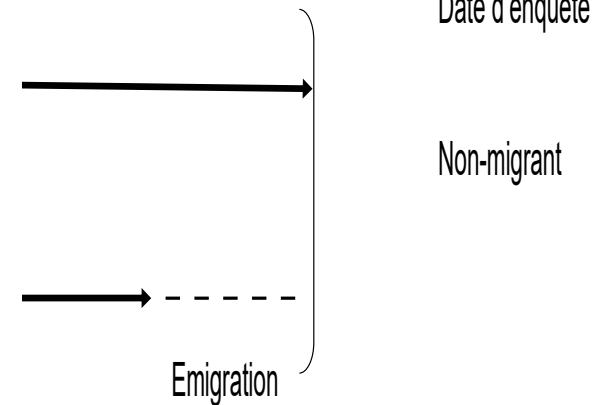
► Durée avant l'événement



Notions d'événement et de troncatures

- ▶ Troncature : c'est une sorte de censure
- ▶ Les censurés : Les exclus vivants, les perdus de vue
- ▶ Il est difficile d'observer les individus sur toute leur vie ou même sur un intervalle de temps donné

- ▶ **Troncature à droite:** La durée est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. L'observation de l'évènement est interrompue par la date de l'enquête ou par une émigration.



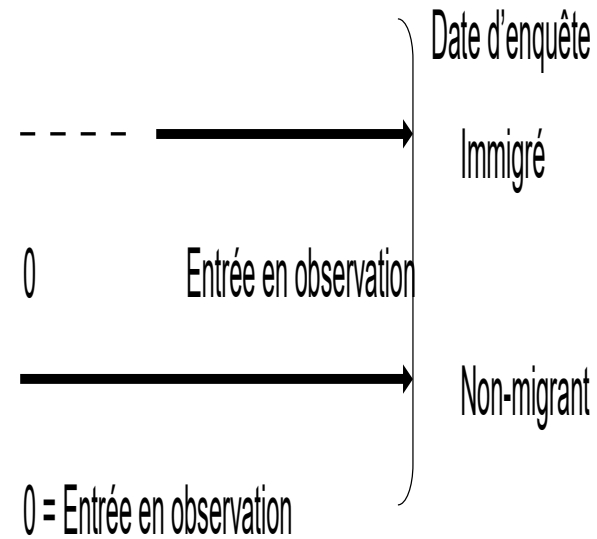
Notions de troncatures

Troncature à gauche : Entrée dans le champ d'observation (Immigration)

Cette notion de données censurées est une caractéristique dont il faudra tenir compte au moment de l'analyse des données de ces études.

Elle définit aussi un biais de censure à gauche ou à droite (left-censorship bias, right-censorship bias) dont il faut être conscient pour interpréter les résultats des études d'incidence.

Il y a aussi d'autres types de censure qui n'interviennent pas dans ce présent cours. C'est celui de censure par intervalle et de censure exacte.



Notions d'événement

Etat initial	Etat final	Evénement	Origine	Durée
Vivant	Mort	Décès	Naissance	Age au décès
Célibataire	Marié	Mariage	Naissance	Age au mariage
Non instruit	instruit	instruction	6ans	Durée de scolarisation



Collecte de données et Outils de collecte



Collecte de données biographies

- ▶ Il est question de retracer les événements connus par une personne à partir d'un moment donné de sa vie jusqu'à la date de l'enquête;
- ▶ Ce type de collecte nécessite de faire appel à la mémoire à court et à moyen terme, de bien repérer dans le temps les événements vécus par les enquêtés.

Outils de collecte

- ▶ Questionnaire biographique et fiche **AGEVEN**

Sources de données biographiques

- ▶ les Enquêtes rétrospectives, les Enquêtes « prospectives » (enquêtes à passages répétés, registres de population, observatoires de population)
- ▶ Qu'en est-il des données que vous pouvez collecter vous-même à un temps donné? (exemple d'enquête permettant d'avoir ces types de données?)
- ▶ Où trouver les données biographiques: dans les instituts de recherche comme ISSP.



Importants

- ▶ Variable à expliquer ou variable dépendante
- ▶ La durée avant l'occurrence d'un événement précis (durée avant la première entrée dans le système éducatif, durée avant l'entrée dans le système éducatifs migration, décès, etc.), mesurée par rapport à une origine.
- ▶ On s'intéresse au risque que l'événement se produise en fonction de la durée écoulée depuis une origine.
- ▶ La durée ou le risque dépend de : Variables explicatives presque fixes (qui ne varient pas comme le sexe, l'ethnie, etc.) et des variables en fonction du temps (état matrimonial, statut d'activité)



Quelques limites

- ▶ Approche basée sur la mémoire de moyen ou long terme
- ▶ Risques d'oubli ou de la datation imprécise de certains événements
- ▶ L'analyse s'intéresse aux groupes d'individus
- ▶ Problème de représentativité de groupe pour certaines transitions (petite taille de l'échantillon)
- ▶ La succession des événements importe plus que leur datation précise

La méthode de Kaplan-Meier

- ▶ La méthode de Kaplan-Meier permet d'estimer les **fonctions de survie** sans que nécessairement les intervalles de temps soient réguliers. Les courbes de survie permettent d'analyser l'évolution de l'effectif d'une population donnée avec le temps.
- ▶ Cette technique est utilisée pour l'analyse de données de survie, qu'il s'agisse d'individus (recherche sur la leucémie par exemple), ou de produits (résistance au temps d'un outil de production par exemple).
- ▶ On entendra parler beaucoup de **sortie d'observation ou d'étude**: l'on sort de l'étude si certains individus meurent (les produits se cassent ou être hors d'usage), **mais** d'autres sortent de l'étude parce qu'ils guérissent, ou encore lorsque l'on perd leur trace (déménagement par exemple) ou parce que l'étude est interrompue. Le premier type d'information est appelé « données événement », tandis que le second est appelé « données censurées ».



La méthode de Kaplan-Meier

L'analyse de Kaplan-Meier permet aussi de comparer des groupes de populations, en s'appuyant sur leur courbe de survie.

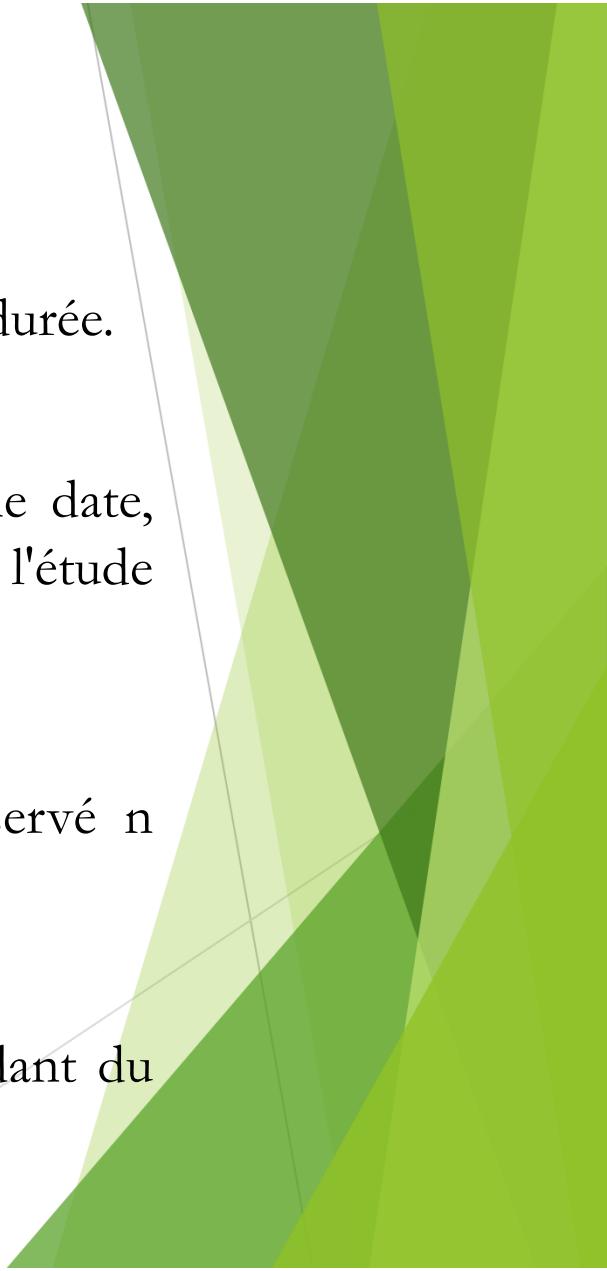
Par exemple, il peut être intéressant de comparer les temps de survie des garçons et des filles dans le supérieur, ou de comparer les temps de casse pour un même produit fabriqué sur deux chaînes de production différentes.

L'utilisation de la méthode Kaplan-Meier implique que l'on fasse l'hypothèse que les observations sont indépendantes.

De même, on fait l'hypothèse que la censure est indépendante : soient deux individus pris au hasard, inclus dans l'étude au temps $t-1$; si l'un d'eux est censuré au temps t , alors leur chance de survie est égale au temps t . On distingue quatre types de censure indépendante :

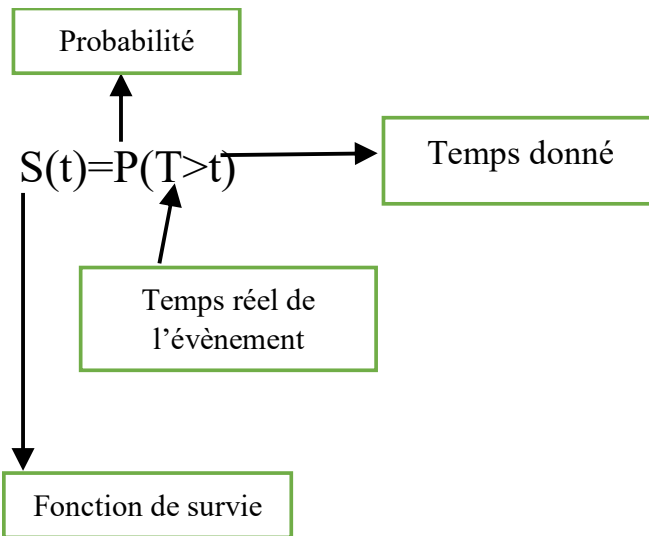


La méthode de Kaplan-Meier

- ▶ **Type I simple** : tous les individus sont censurés après une même durée.
 - ▶ **Type I progressif** : tous les individus sont censurés à la même date, quelle que soit la durée pendant laquelle ils ont été suivis (fin de l'étude par exemple).
 - ▶ **Type II** : les individus sont suivis jusqu'à ce que l'on ait observé n événements.
 - ▶ **Aléatoire** : le temps auquel se produit une censure est indépendant du temps de survie
- 


La fonction de survie: un élément de base

- ▶ **La fonction de survie:** C'est la probabilité qu'un patient vive au moins un certain temps après le diagnostic de sa maladie (peut-être un cancer), elle prédit le pronostic favorable ou défavorable et cela peut permettre au médecin de mieux adapter le traitement pour un patient.
- ▶ Exemple: le temps que met un malade avant de guérir du cancer





Méthode d'analyse descriptive : estimateurs de Kaplan-Meier

- ▶ Estimation à partir de données observées
 - ▶ Observation des temps d'événement
 - ▶ Prendre en compte les données censurées
 - ▶ Principe : à chaque temps d'événement
 - ▶ – Identifier les sujets à risque
 - ▶ – Compter les événements
- 

Méthode d'analyse descriptive : estimateurs de Kaplan-Meier

Tables de séjour de Kaplan-Meier

Elles permettent de décrire les événements qui varient au cours du temps.

Estimateurs de Kaplan-Meier

Ils mesurent la probabilité d'occurrence de l'événement à un instant t_i pour les individus n'ayant pas jusque-là connu l'événement étudié.

Courbe de survie de Kaplan-Meier

- ▶ Elle représente la distribution de la durée avant la réalisation de l'événement.
- ▶ Elle s'interprète comme la proportion de « survivant » pour chaque durée de séjour dans un état donné.

Tables de séjour de Kaplan-Meier

Quotients instantanés $h(t_i) = \frac{d(t_i)}{N(t_i)}$

$d(t_i)$ = nombre d'événements observés en t_i ; $N(t_i)$ = effectif soumis au risque juste avant t_i ; $t_1, t_2, \dots, t_k, \dots, t_n$, âges auxquels les individus ont connu l'événement étudié

Proportion des survivants à chaque temps (âge) :

$$S(t_i) = \prod_{j \leq i} (1 - h(t_j))$$

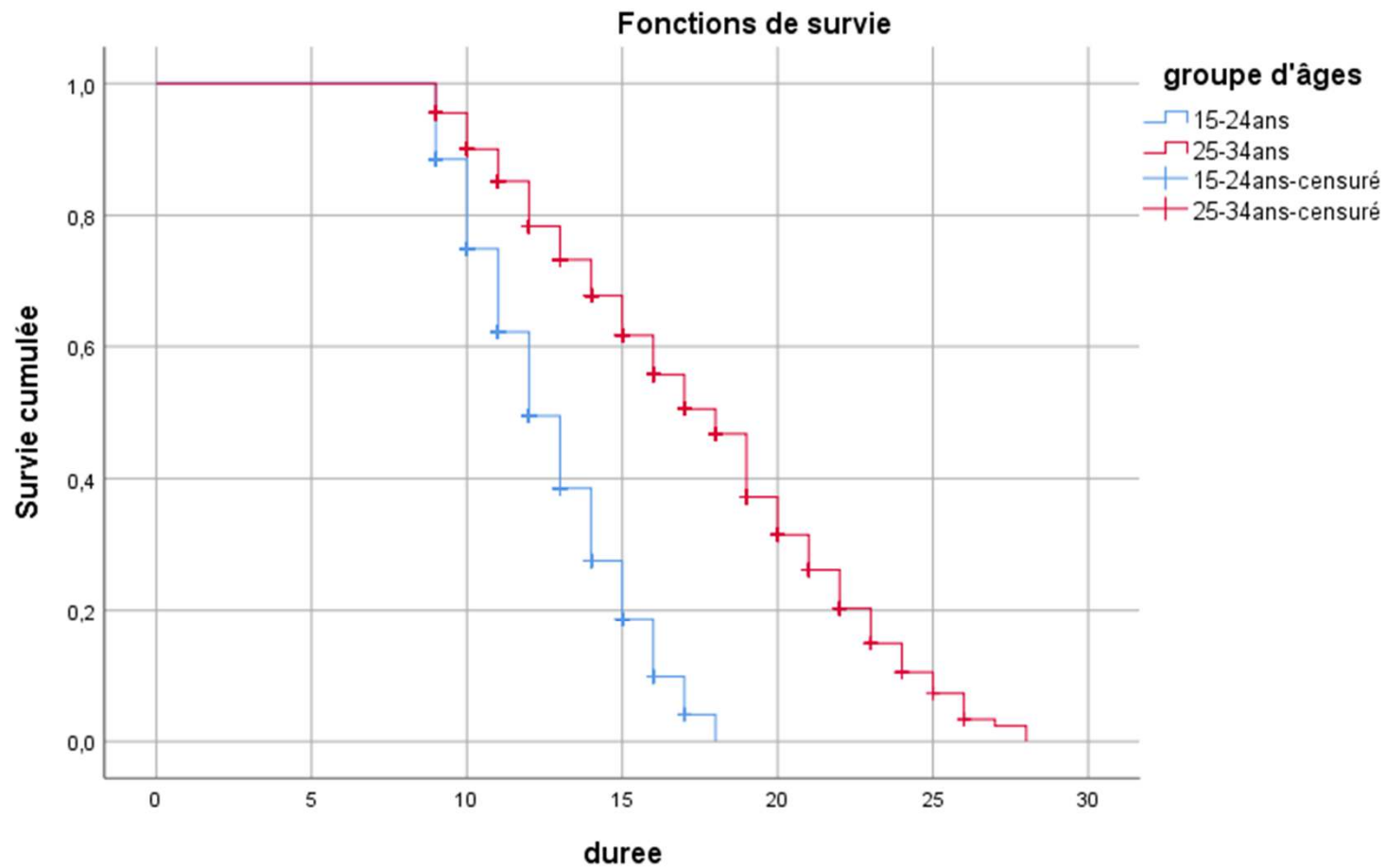
$$h(t_1) = \frac{3}{50} = 0,06 \Rightarrow S(t_1) = 1 - 0,06 = 0,94$$

$$h(t_2) = \frac{0}{47} = 0 \Rightarrow S(t_2) = [1 - h(t_1)][1 - h(t_2)] = 0,94$$

Exemple d'une table de séjour de Kaplan-Meier

Intervalle de temps (t_i)	Nombre d'individus en début d'intervalle $N(t_i)$	Evénements $d(t_i)$	Troncatures	Quotient s $h(t_i)$	Proportion des survivants $S(t_i)$	
1	2	50	3	0	0,060	0,94
2	3	47	0	1	0,000	0,94
3	4	46	3	2	0,065	0,88
4	5	41	2	0	0,049	0,84
5	6	39	1	2	0,026	0,82

Courbe de survie, Kaplan Meier



H1: La distribution des durées de survies est la même quel que soit la cohorte

H0: il y a une différence de distribution des durées de survies entre les cohortes
Conclusion: accepter H0

Le modèle semi paramétrique de regression de cox

- effectuer une modélisation et de mesurer l'impact de variables explicatives, presque fixes ou dépendantes du temps, sur le quotient instantané.
- repose sur l'analyse de risque que l'événement étudié se produise en fonction de la durée écoulée depuis une origine.
- combinaison de deux types d'analyse : la régression et les tables de séjours.
-
- Ainsi, le modèle de Cox résulte de l'introduction du facteur temps dans la régression. Pour cela, il est mieux adapté à l'étude des phénomènes qui évoluent dans le temps.

Méthode d'analyse explicative : régression semi-paramétrique de Cox

- ▶ La méthode de Cox permet de mesurer le risque relatif instantané de connaître l'événement étudié si on a une caractéristique z_i :

$$h(t, z) = h_0(t) \exp(z\beta)$$

- ▶ $h_0(t)$ est le quotient instantané pour la catégorie de référence ($z = 0$)
- ▶ β représente les coefficients estimés de z_i :

$$z\beta = z_1\beta_1 + z_2\beta_2 + \dots + z_n\beta_n$$

- ▶ L'exponentiel de chaque coefficient ($z\beta$) fournit le risque relatif de connaître l'événement par rapport à la catégorie de référence

$$\exp(\beta_1) = \frac{h(t, z_1)}{h(t, 0)}$$

Méthode d'analyse explicative : régression semi-paramétrique de Cox

- ▶ L'effet des variables est proportionnel à la probabilité annuelle de connaître l'événement étudié. Le modèle de cox est aussi appelé modèle à risques proportionnels.
- ▶ La regression de cox est dite semi-paramétrique car dans son modèle, elle a une composante non paramétrique qui est la fonction de séjour de base ($h_0(t)$) et une autre composante paramétrique, qui est le vecteur de variables explicatives $\exp(z,\beta)$ d'où la formulation mathématique $h(t,z) = (h_0(t)) \exp(z,\beta)$.

Régression semi paramétrique de cox: un exemple de résultat sur SPSS

Variables dans l'équation

	B	Erreur standard	Wald	df	Sig.	Exp(B)	95,0% CI pour Exp(B)	
							Plus bas	Supérieur
quel est/était votre statut d'occupation dans ce logement à			117,573	2	,000			
quel est/était votre statut d'occupation dans ce logement à (1)	-2,208	,450	24,065	1	,000	,110	,046	,266
quel est/était votre statut d'occupation dans ce logement à (2)	-2,383	,245	94,689	1	,000	,092	,057	,149
sexe	-,039	,050	,594	1	,441	,962	,871	1,062
quelle est votre religion à la naissance?			8,411	4	,078			
quelle est votre religion à la naissance?(1)	-,312	,157	3,948	1	,047	,732	,538	,996
quelle est votre religion à la naissance?(2)	-,379	,159	5,678	1	,017	,685	,502	,935
quelle est votre religion à la naissance?(3)	-,234	,169	1,928	1	,165	,791	,569	1,101
quelle est votre religion à la naissance?(4)	-,431	,201	4,589	1	,032	,650	,438	,964
âge de l'enquêté	-,172	,007	571,583	1	,000	,842	,831	,854


H1: il y a une différence entre ceux qui sont sortis (1^{ère} fois) à 15-24ans et ceux sortis à 30-34 ans du toit parental

H0: il n'y a pas de différence entre ceux qui sont sortis à 15-24ans (1^{ère} fois) et ceux sortis à 30-34 ans



Exemple pratique : Première sortie du toit parental sur SPSS

Étapes de l'analyse

1. Description des données biographiques et opérationnaliser
 2. Conceptualisation d'un événement
 3. Construction du fichier d'analyse
 4. Analyse des données
- 

Description de l'événement : Première sortie de l'école

- ▶ Lieu: à Bafia
- ▶ Année de l'enquête : 2010
- ▶ Population cible : Individus âgés de 15 à 34 ans en 2010
- ▶ Strates retenues : Hommes et femmes appartenant aux groupes d'âges 15-24 ans et 25-34 ans
- ▶ Taille de l'échantillon : 4949

Description et Opérationnalisation de l'événement : Première sortie de l'école

- ▶ **Evénement étudié:** Première sortie du toit parental
- ▶ **Origine de l'observation:** ceux qui vivent sous le toit parental de 6 ans
- ▶ **Population soumise au risque:** Tous les individus vivant avec leur parents
- ▶ **Durée d'observation:** âge au moment de la première sortie du toit– âge depuis 6 ans
- ▶ **Troncatures:** Ceux qui fréquentent encore l'école au moment de l'enquête ;
Durée = date de l'enquête – date d'entrée à l'école

Construction du fichier d'analyse

- ▶ **Indice de troncature** qui prend la valeur 1 si l'événement se réalise et 0 sinon
- ▶ Le **début** d'observation doit être un instant de référence commun à tous les individus. Donc la date de début d'observation = $dnais + 6$
- ▶ La date de fin d'une période est égale à la date de début de la période suivante.

Pour les personnes sortie de l'école pour la première fois, cela correspond à la date de sortie. Pour les censurés (perdus de vue) cela correspond à la date de censure. Pour les personnes tronquée à la date de l'enquête cela correspond à la date de l'enquête.



Construction du fichier d'analyse

► **Durée d'observation** (avant l'événement s'il se réalise)

Pour ceux qui sortent de l'école avant la date de l'enquête, il s'agit du temps écoulé depuis l'âge de 6 ans jusqu'à la date de première sortie de l'école.

Pour ceux qui ne sortent pas de l'école, il s'agit du temps écoulé depuis l'âge de 6 ans jusqu'à la date de l'enquête.

Attention: Ne pas oublier de trier les individus par identifiant!!



Conceptualisation d'un événement

- ▶ **Deux fichiers de données sont fournis** (aller à vue des données)

Un fichier de données sur les caractéristiques sociodémographiques des enquêtés (**indiv.sav**)

Un fichier de données événementielles sur les différentes périodes de logements occupés par les enquêtés (**base_loge.sav**)

- ▶ Dans le premier fichier. Chaque ligne correspond à un individu et chaque colonne à une variable
- ▶ Le second fichier est un fichier biographique. Chaque ligne correspond à un événement (transition) et chaque colonne une variable
- ▶ Fusion des deux bases et **ne grader que les fusions parfaites:**

merge 1:m ident using

```
"C:\Users\USER\Dropbox\PC\Desktop\COURS_2021_2022\MASTER_II_RECHERCHE_2021_2022\
Master_II_Recherc_SPSS_2021_2022\Biographie\Bases\pour \bbase_loge.dta ", keep(1 2 3)
```

Chemin d'accès dépend de
l'ordinateur

Nom de la base à
merger

Conceptualisation d'un événement

- ▶ Nécessité de fusionner les deux fichiers afin de rassembler les informations d'où l'importance d'une fusion des deux fichiers et **ne grader que les fusions parfaites**
- ▶ **NB:** Pour faire l'analyse des biographies, on a besoin de deux éléments fondamentaux

Durée d'observation (avant l'événement s'il se réalise)

L'indice de troncature qui prend la valeur 1 si l'événement se réalise et 0 sinon

Le début d'observation doit être un instant de référence commun à tous les individus

La date de fin d'une période est égale à la date de début de la période suivante

Résumé

► Un fichier biographique

il retrace l'histoire de vie d'un individu, les événements survenus dans la vie de l'individu depuis sa naissance jusqu'à la date de l'enquête.

► Est-ce possible que l'identifiant d'un individu se répète dans un fichier biographique?

Oui, dans un fichier biographique chaque ligne correspond à un événement survenu dans la vie de l'individu (passage d'un état qualitatif à un autre). Donc lorsque l'identifiant d'un individu se répète, cela signifie qu'il y a eu un changement d'un état à un autre.

Résumé

- ▶ **Troncature à droite** : C'est une interruption de l'observation de l'individu alors qu'il n'a pas encore connu (subi) l'événement étudié. En effet, lorsqu'un individu n'a pas encore subi l'événement étudié et qu'il sort du champ d'observation, il est tronqué à droite. (On dit droite parce que l'interruption a lieu à droite de l'échelle du temps qui par convention va de la gauche vers la droite).
- ▶ **Exemple de troncature à droite : la date de l'enquête.**
- ▶ **Troncature à gauche** : c'est l'entrée de l'individu dans le champ de l'observation au cours de l'intervalle considéré.

Exemple de troncature à gauche : l'immigration.

- ▶ En effet, avant d'immigrer l'individu n'est pas soumis au risque de subir l'événement dans la population étudiée. Il n'est soumis à ce risque qu'à partir du moment de son immigration dans la population étudiée

Résumé

- ▶ Les deux éléments essentiels dans une analyse des biographies :
- ▶ **L'indice de troncature** relatif à l'événement étudié
- ▶ **La durée d'observation avant l'occurrence de cet événement.** (Le temps écoulé depuis une origine jusqu'à la sortie du champ d'observation (la variable indicatrice donnera la durée jusqu'à l'événement si on a subi l'événement, jusqu'à la date de l'enquête si on ne l'a pas connu, et des missing values si est tronqué à droite par la date de l'enquête)

Résumé

► Les deux principaux outils d'analyse des biographies.

■ La table de survie (méthode descriptive)

Le principe de base de la table de survie est de considérer le temps qui s'écoule entre un instant de référence commun à tous les individus soumis au risque et la date de l'événement ou de sortie d'observation. On découpe ensuite le temps en intervalle (annuel, mensuel, etc.), ce qui permet de calculer des probabilités (des risques) que l'évènement se produise dans les intervalles parmi les individus soumis au risque.

La table de survie permet donc de passer de durée observées à des risques de connaître l'événement au cours du temps, elle tient compte des troncatures à droite.

- Le modèle de régression de cox (**méthode explicative**): Il est une combinaison de deux types d'analyse : la régression et les tables de séjours. Ainsi, le modèle de Cox résulte de l'introduction du facteur temps dans la régression. Pour cela, il est mieux adapté à l'étude des phénomènes qui évoluent dans le temps.

Résumé

- ▶ **Utilité de l'indice de troncature dans une analyse biographique.**
- ▶ L'indice de troncature permet de déterminer :
- ▶ si l'individu subi l'événement étudié : il sort du champ d'observation. (cet indice prend la valeur 1 dans ce cas,
- ▶ ou alors s'il ne subi pas l'événement étudié, il reste dans le champ d'observation (dans ce cas l'indice prend la valeur 0)
- ▶ ou alors s'il ne nous intéresse plus : parce qu'il est déjà sorti du champ d'observation et il n'est donc plus soumis au risque de subir l'événement étudié. (dans ce cas il prend la valeur : .).
- ▶ C'est grâce à cet indice qu'on effectuera les analyses

Résumé

Décrire brièvement les principaux outils d'analyse des biographies ?

- ▶ La table de survie de Kaplan-Meier,
- ▶ la régression semi-paramétrique de Cox.
- ▶ Le premier est un tableau descriptif des intervalles de temps de référence pour l'événement étudié, les individus exposés à chaque période et des probabilités d'échapper à l'événement à chaque temps.

Résumé

- ▶ Les deux principaux outils d'analyse des biographies:

La table de survie (méthode descriptive)

- ▶ Le principe de base de la table de survie est de considérer le temps qui s'écoule entre un instant de référence commun à tous les individus soumis au risque et la date de l'événement ou de sortie d'observation. Elle permet de passer de durée observées à des risques de connaître l'événement au cours du temps.

Résumé

Le modèle de régression de cox (méthode explicative): Il est une combinaison de deux types d'analyse : la régression et les tables de séjours. Ainsi, le modèle de Cox résulte de l'introduction du facteur temps dans la régression. Pour cela, il est mieux adapté à l'étude des phénomènes qui évoluent dans le temps.

Rôle de l'analyse des biographies: Retracer et expliquer l'évolution d'un phénomène et d'en dégager les facteurs explicatifs.

- ▶ Analyser les événements en interaction les uns par rapport aux autres
- ▶ Calculer les risques sur des populations homogènes

Références Bibliographiques

- ▶ Zakaria RAITI (2017), Modélisation de la durée de maintien en arrêt de travail d'une population de travailleurs non-salariés, mémoire de statisticien, Institut des Actuares, 142 p.
- ▶ Nganawara D. (2016-2017) Cours de biographie, Analyse des biographies avec le logiciel stata, Master I, IFORD
- ▶ COURGEAU Daniel et LELIEVRE Evra (1989), Analyse démographique des biographies, INED, 268p.
- ▶ COURGEAU Daniel et LELIEVRE Evra (1995), L'apprentissage de l'analyse biographique en démographie, in, L'analyse longitudinale en économie sociale, CNRS
- ▶ <https://www.xlstat.com/fr/solutions/fonctionnalites/analyse-kaplan-meier>
- ▶ https://fad.univ-lorraine.fr/pluginfile.php/23863/mod_resource/content/2/co/Notion_Donnees_Censurees.html

BONNE LECTURE

