INTRODUCTION À L'ANALYSE MULTIPLE ET L'ANALYSE DE SURVIE

UE MRCE

28/03/2024

Léa Duchesne

Contact : lea.duchesne@univ-nantes.fr

Quel modèle pour quelle question ?

« Tous les modèles sont faux, mais certains sont utiles »

George Box

- Ce que l'on va voir dans ce cours : **modèles de régression**
- **Modélisation mathématique** du phénomène étudié à partir des données d'un échantillon

  - Variable **dépendante** (Y) – « à expliquer »

    → Quantitative ou qualitative

  - Variables **indépendantes** (X1, X2, …, Xn) – « explicatives »

    → Exposition principale

    → Autres facteurs de risque (potentiels facteurs de confusion)

    → Quantitatives ou qualitatives

$$Y = f(X)$$

Quel modèle pour quelle question ?

- Quelle **fonction** (mathématique) choisir ?

  → Dépend de :

  - La nature de la variable à expliquer +++
  - Conditions de validité des modèles (distribution des variables, etc.)

Quel modèle pour quelle question ?

- Si Y = variable quantitative

  ⟶ Régression linéaire

  Espérance (= moyenne) de Y sachant $X_1, X_2, \ldots, X_n$

- Si Y = variable qualitative binaire

  ⟶ Régression logistique

  Risque de survenue de Y sachant $X_1, X_2, \ldots, X_n$ dans une période fixée

- Si Y = variable qualitative binaire + notion de délai

  ⟶ Analyse de survie (Kaplan-Meier, modèle de Cox, …)

  Incidence instantanée de Y sachant $X_1, X_2, \ldots, X_n$

RÉGRESSION LINÉAIRE SIMPLE

Principe

Quantitative

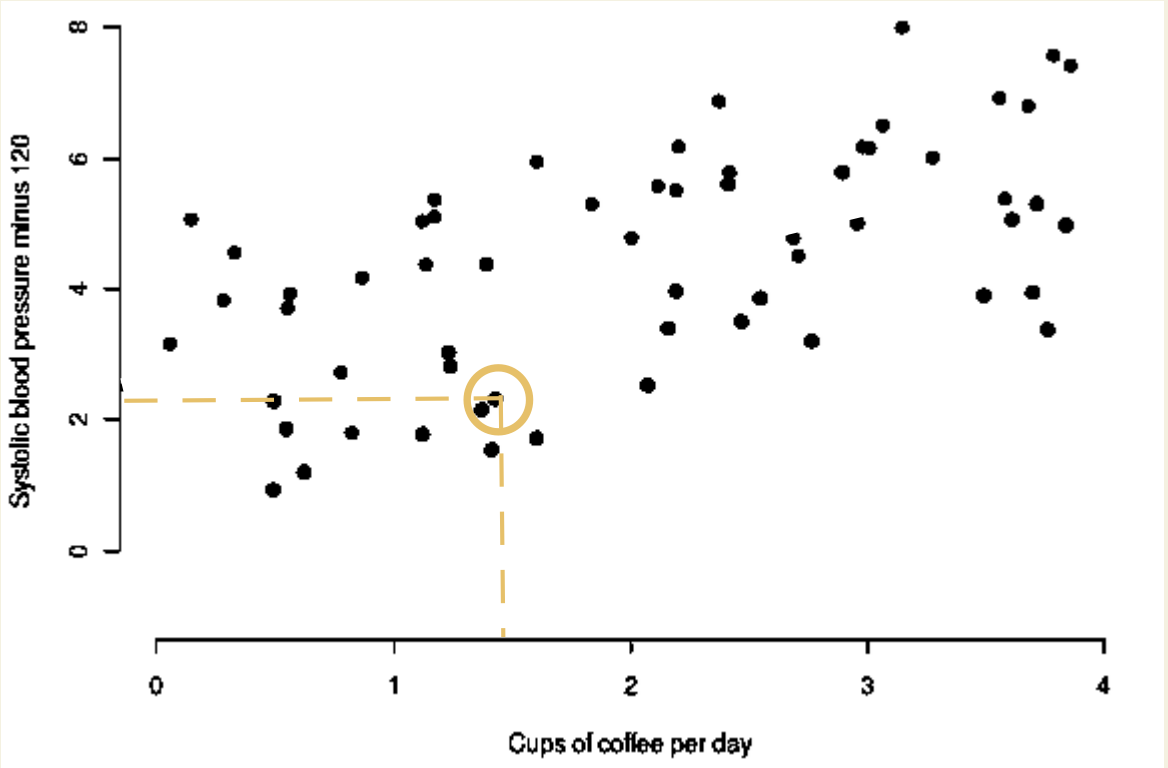
Variable **dépendante** (Y)

Quantitative ou qualitative

Variable **indépendante** (Y)

Exemple : étude du lien entre la pression systolique et la consommation de café à partir des données d'un échantillon de personnes issues de la population générale.

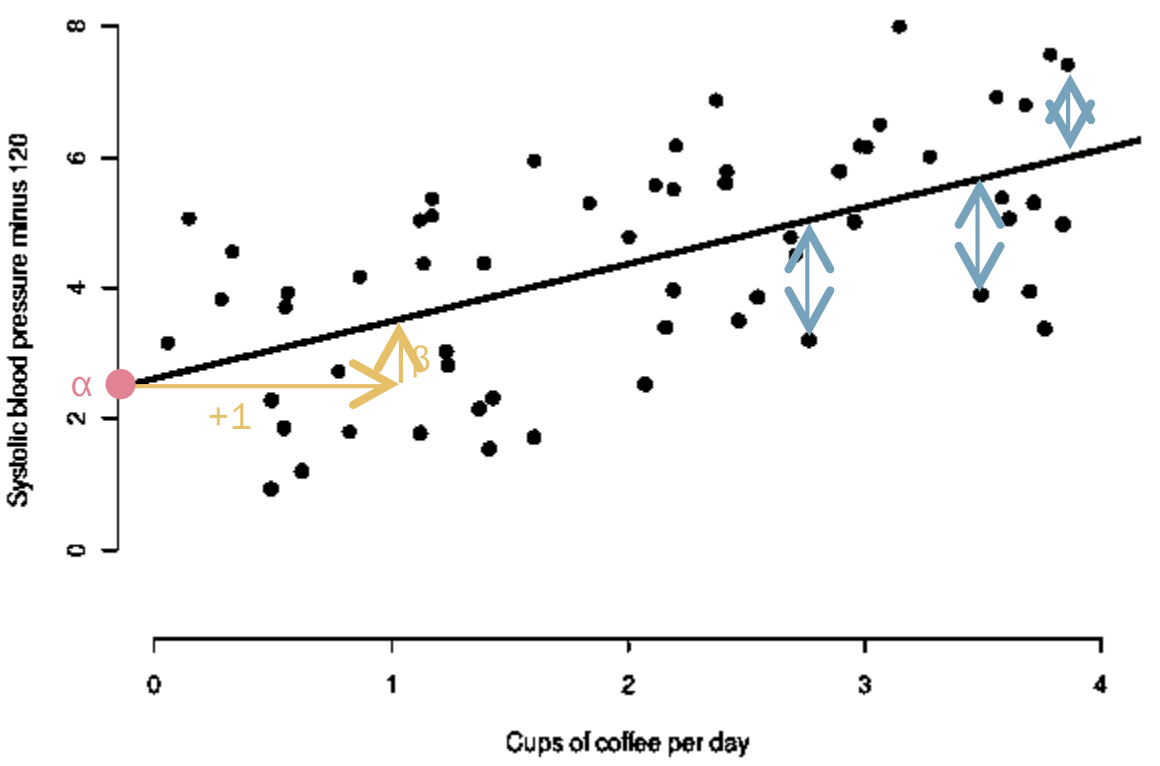
Id	Pression systolique	Nombre de tasses de café par jour
1	2,2	1,4
2	6,0	1,7
…	…	…



Principe

Quelle fonction ?

→ Fonction affine $y = \alpha + \beta x$



Pour un individu i, la valeur de Y est égale à :

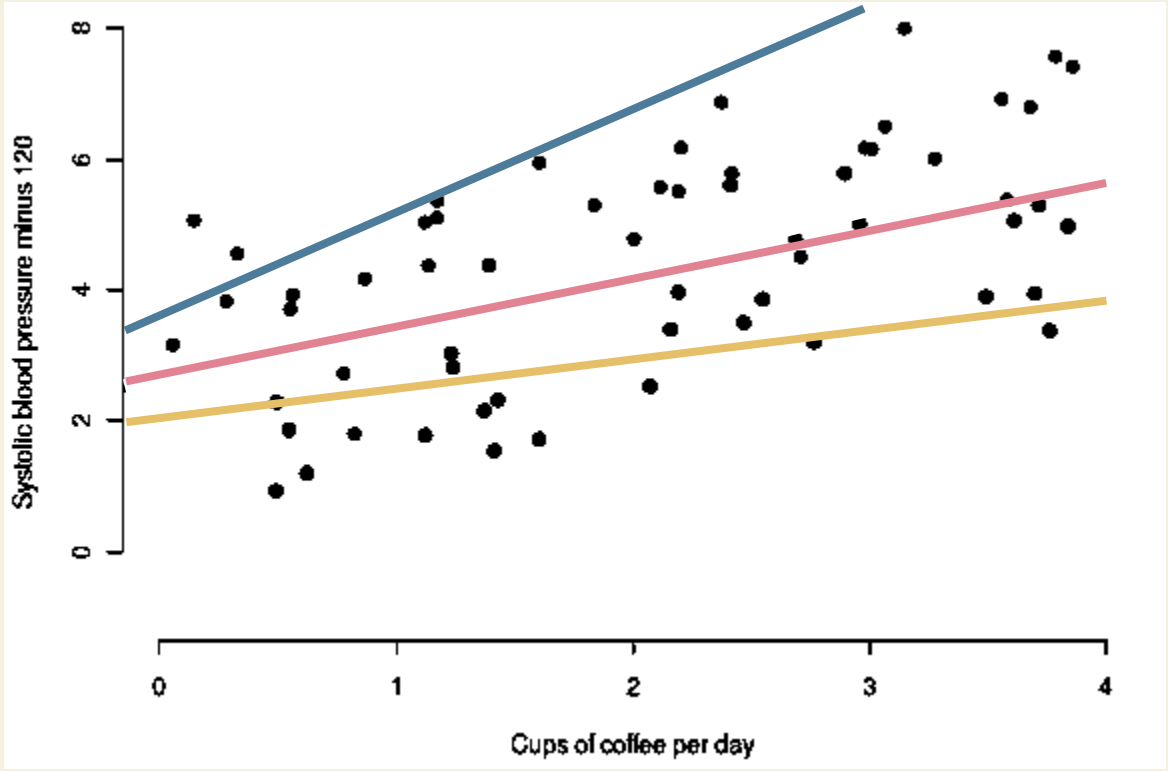
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- β_0 : Ordonnée à l'origine (*intercept*)
- β_1 : Pente (*slope*)
- ε_i : Erreurs

Estimation des paramètres

<u>Quelles valeurs de α et de β ?</u>

→ Droite qui représente le mieux les données



Estimation des paramètres

Quelles valeurs de α et de β ?

→ Droite qui représente le mieux les données

→ Droite qui minimise les **résidus** (ε) : méthode des **moindres carrés** (Ordinary Least Squares)

- **Somme des carrés des écarts** (SCE)

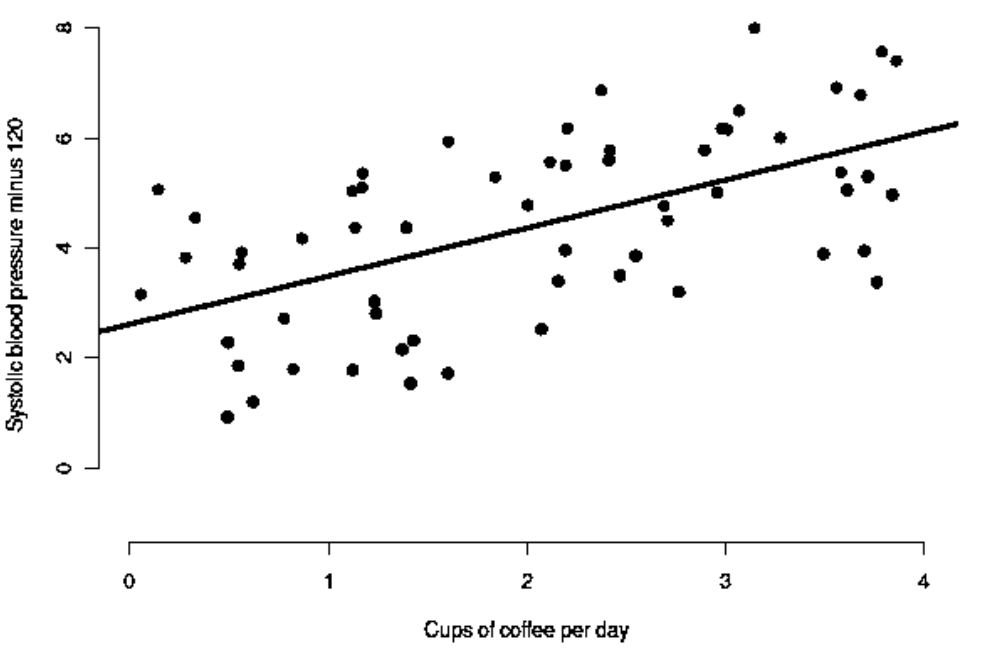
$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \underbrace{a - b.x_i}_{})^2$$

- Pour chaque droite $y = \alpha + \beta.x$ possible, c.à.d. pour chaque couple de valeurs (α, β)

  → SCE différente

  → **On choisit la droite qui minimise la SCE**

Interprétation : β_0



β_0 : moyenne de Y chez les personnes pour qui $X_1 = 0$

→ Attention : pas de sens pour certaines variables, dans ce cas ne pas l'interpréter !

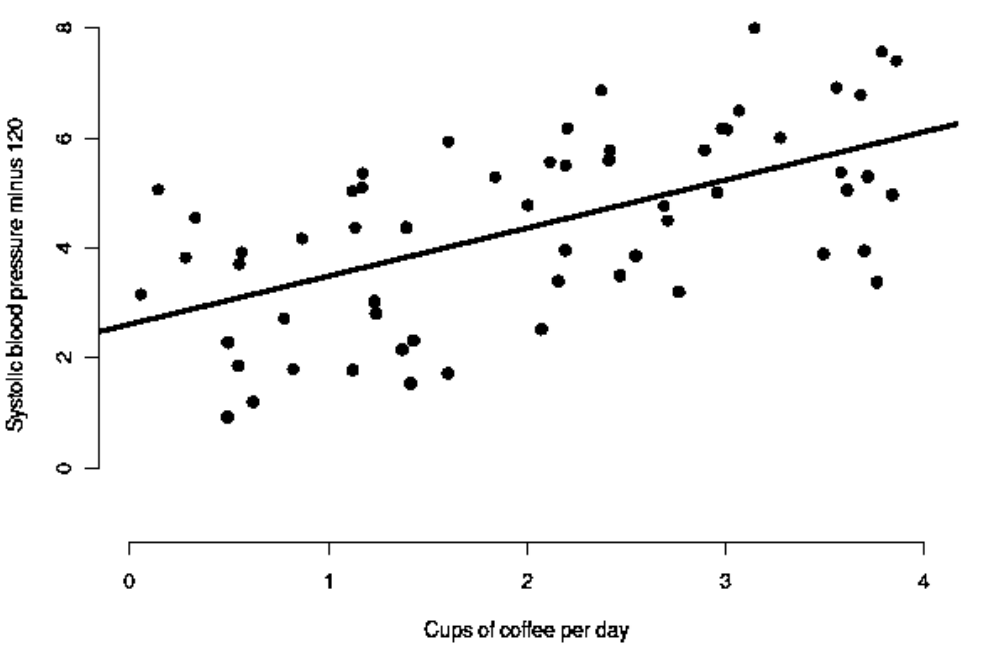
Exemple : si X = âge

Variable	Coefficient	p-value
Intercept	2,61	$2x10^{-9}$
Consommation de café	0,87	$7x10^{-7}$

En moyenne, la pression systolique des personnes ne consommant pas de café est de 2,61

Interprétation : β_1

Interprétation de β_1 dépend de la nature de la variable



Cas n°1

Ici X_1: quantitative

→ β_1 = évolution moyenne de Y lorsque X_1 augmente d'une unité

- $\beta_1 > 0$: si X_1 augmente, Y augmente
- $\beta_1 < 0$: si X_1 augmente, Y baisse

Variable	Coefficient	p-value
Intercept	2,61	$2x10^{-9}$
Consommation de café	0,87	$7x10^{-7}$

Pour une tasse de café supplémentaire, la pression systolique augmente en moyenne de 0,87

Interprétation : β_1

Modalité (« catégorie », « classe ») de référence : bien la repérer pour interpréter les résultats !

Variable	Coefficient	p-value
Intercept	2,61	$2x10^{-9}$
Consommation de café (oui/non)	0,87	$7x10^{-7}$

Cas n°2

Si X_1: qualitative binaire

→ β_1 = écart entre la moyenne de Y chez les personnes chez qui $X_1 = 1$ et celles chez qui $X_1 = 0$

Les personnes consommant du café ont une pression systolique plus élevée de 0,87 en moyenne que les personnes ne buvant pas de café

Problème : codage peu précis

→ Boire du café augmente de β_1 unités le risque, quelque soit la dose de café consommée

Cas n°3

On pourrait aussi inclure cette variable sous forme **catégorielle**, par exemple :

- 0 tasses
- Entre 1 et 5 tasses par jour
- > 5 tasses par jour

Inclure une variable qualitative catégorielle nécessite de réaliser un **modèle de régression linéaire multiple**

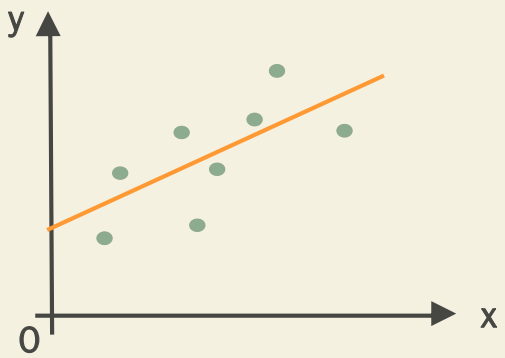
Existe-t-il un lien linéaire entre Y et X ?

Test sur le coefficient β_1

→ β_1 : différent de 0 ?

Lien linéaire (positif ou négatif)

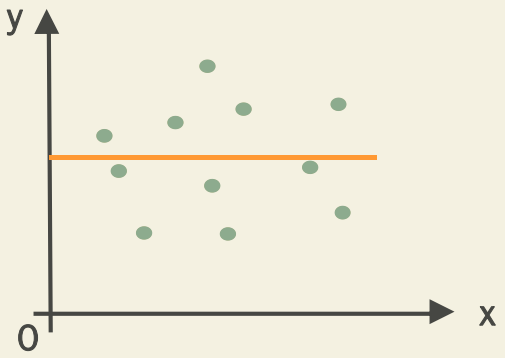
Pente > 0



Pente < 0

Absence de lien

Pente = 0



Existe-t-il un lien linéaire entre Y et X ?

- Hypothèses du test :

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

- p-valeur associée au coefficient < risque 1ère espèce fixé (souvent 5%)

  → Au risque α , la variable X est liée linéairement à Y

- p-valeur associée au coefficient > risque 1ère espèce fixé

  → Les données (et donc le modèle obtenu) ne permettent pas de montrer que la variable X est liée linéairement à Y

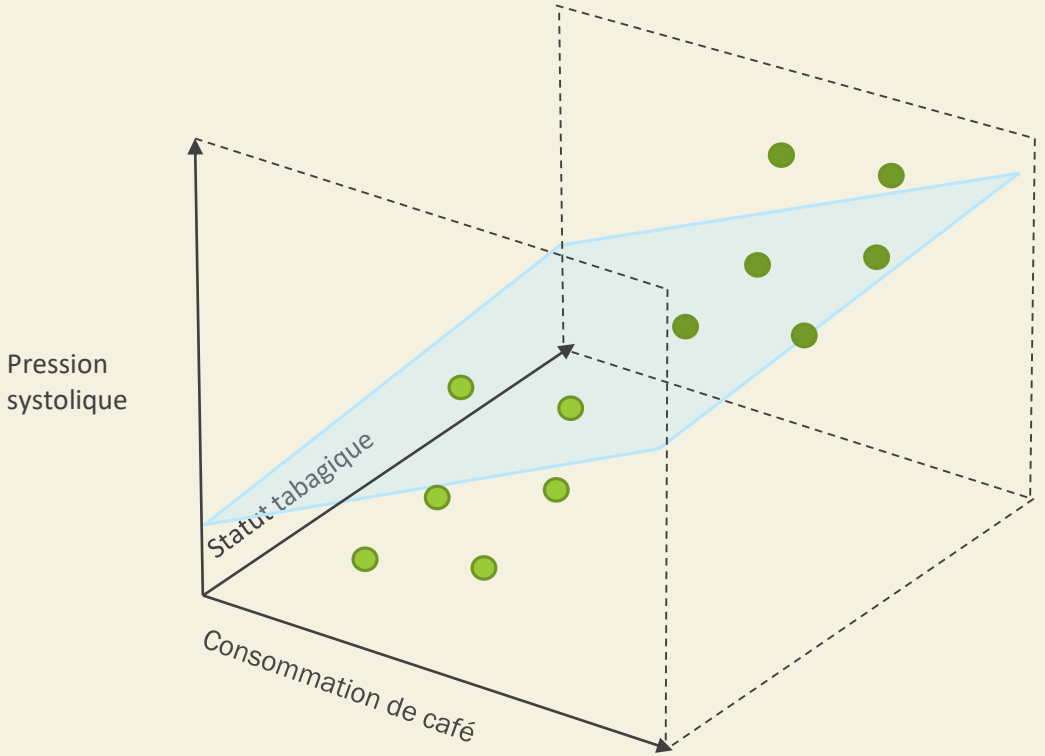
RÉGRESSION LINÉAIRE MULTIPLE

Principe

- Souvent 1 seule variable explicative ne suffit pas
  → Autre variable explicative ? (tabagisme par exemple)

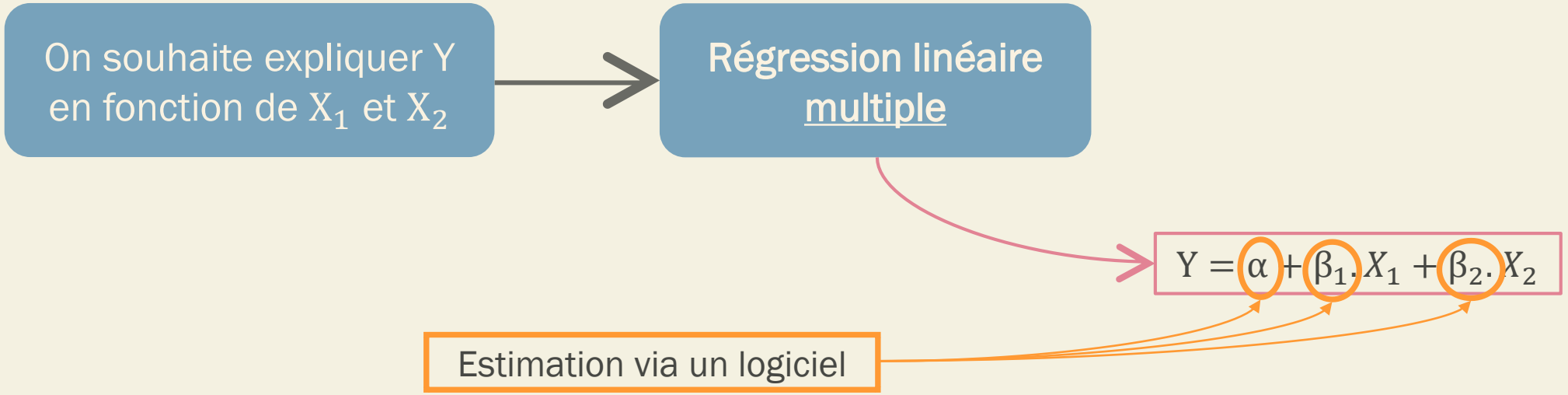
Pour un individu i : $y_i = \alpha + \beta_1.x_{1i} + \beta_2.x_{2i} + \varepsilon_i$

Estimation des paramètres

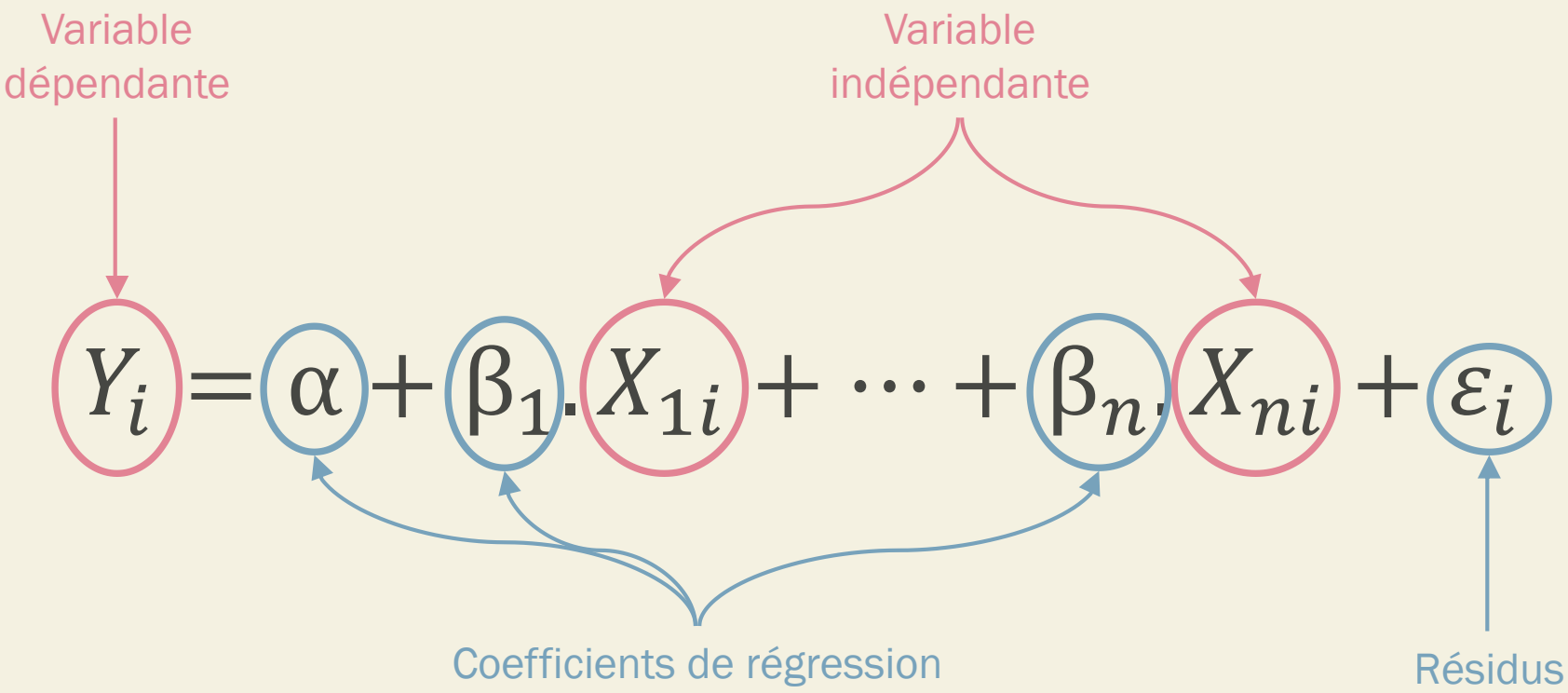


Estimation des paramètres

- Souvent 1 seule variable explicative ne suffit pas
  → Autre variable explicative ? (tabagisme par exemple)



Résumé


Variable dépendante
Variable indépendante
$Y_i = \alpha + \beta_1 . X_{1i} + \cdots + \beta_n . X_{ni} + \varepsilon_i$
Coefficients de régression
Résidus

Interprétation des coefficients

β_0	Moyenne de Y chez les personnes $X_1 = 0$, $X_2 = 0$, etc.
β_1	
X_1 = quantitative	Evolution moyenne de Y lorsque X_1 augmente d'une unité, toutes les autres variables étant égales par ailleurs
X_1 = qualitative binaire	Ecart entre moyenne de Y chez les personnes $X_1 = 1$ et celles $X_1 = 0$, toutes les autres variables étant égales par ailleurs

Cas n°1

Cas n°2

Variable	Coefficient	p-value
Intercept	1,30	$1x10^{-5}$
Consommation de café	0,94	$5x10^{-13}$
Statut tabagique	2,14	$7x10^{-13}$

En moyenne, la pression systolique des personnes ne consommant pas de café et ne fumant pas de tabac est de 1,30

Interprétation des coefficients

β_0	Moyenne de Y chez les personnes $X_1 = 0$, $X_2 = 0$, etc.
β_1	
X_1 = quantitative	Evolution moyenne de Y lorsque X_1 augmente d'une unité, toutes les autres variables étant égales par ailleurs
X_1 = qualitative binaire	Ecart entre moyenne de Y chez les personnes $X_1 = 1$ et celles $X_1 = 0$, toutes les autres variables étant égales par ailleurs

Cas n°1

Cas n°2

Variable	Coefficient	p-value
Intercept	1,30	$1x10^{-5}$
Consommation de café	0,94	$5x10^{-13}$
Statut tabagique	2,14	$7x10^{-13}$

A statut tabagique égal, les personnes consommant a+1 tasses de café par jour ont une pression systolique supérieure de 0,94 en moyenne que les personnes en consommant a tasse.

Interprétation des coefficients

β_0	Moyenne de Y chez les personnes $X_1 = 0$, $X_2 = 0$, etc.
β_1	
X_1 = quantitative	Evolution moyenne de Y lorsque X_1 augmente d'une unité, toutes les autres variables étant égales par ailleurs
X_1 = qualitative binaire	Ecart entre moyenne de Y chez les personnes $X_1 = 1$ et celles $X_1 = 0$, toutes les autres variables étant égales par ailleurs

Cas n°1

Cas n°2

Variable	Coefficient	p-value
Intercept	1,30	$1x10^{-5}$
Consommation de café	0,94	$5x10^{-13}$
Statut tabagique	2,14	$7x10^{-13}$

A nombre de tasse de café consommées égal, les personnes qui fument ont une pression systolique plus élevée que les personnes ne fumant pas de 2,14 en moyenne

Interprétation des coefficients

Si X : variable qualitative catégorielle (> 2 catégories)

→ Besoin d'être recodée pour être incluse dans le modèle

Pour une variable à k modalités → création de k variables binaires dites variables indicatrices (*dummy variables*)

Choix d'une modalité de référence

X_2 à 3 modalités :
- Non fumeur
- Fumeur occasionnel
- Fumeur journalier

X_2 et X_3 à 2 modalités :
- X_1: fumeur occasionnel (X_1 = 1 si oui, sinon = 0)
- X_2: fumeur journalier (X_2 = 1 si oui, sinon = 0)

Cas n°3

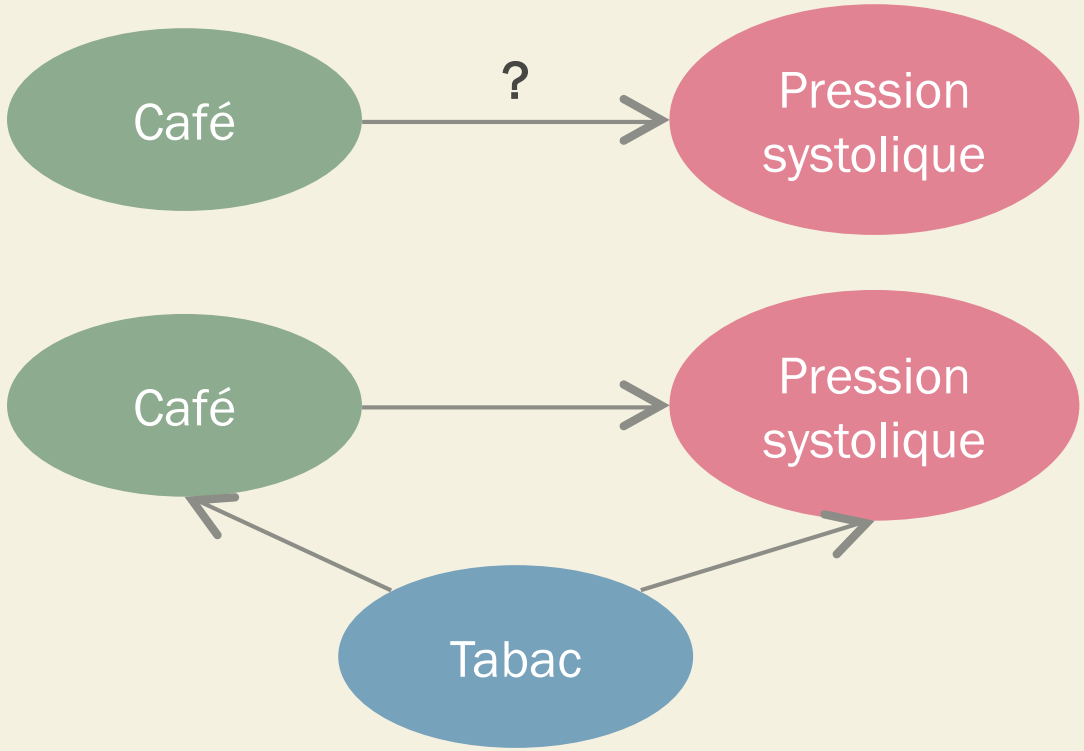
→ β_1: écart entre moyenne de Y chez les personnes X=1 et celles X=0

→ β_2: écart entre moyenne de Y chez les personnes X=2 et celles X=0

Variable	Coefficient	p-value
Intercept	1,30	$1x10^{-5}$
Consommation de café	0,94	$5x10^{-13}$
Statut tabagique		
Fumeur occasionnel	1,43	$7x10^{-12}$
Fumeur journalier	2,30	$6x10^{-12}$

Régression multiple : pourquoi ?

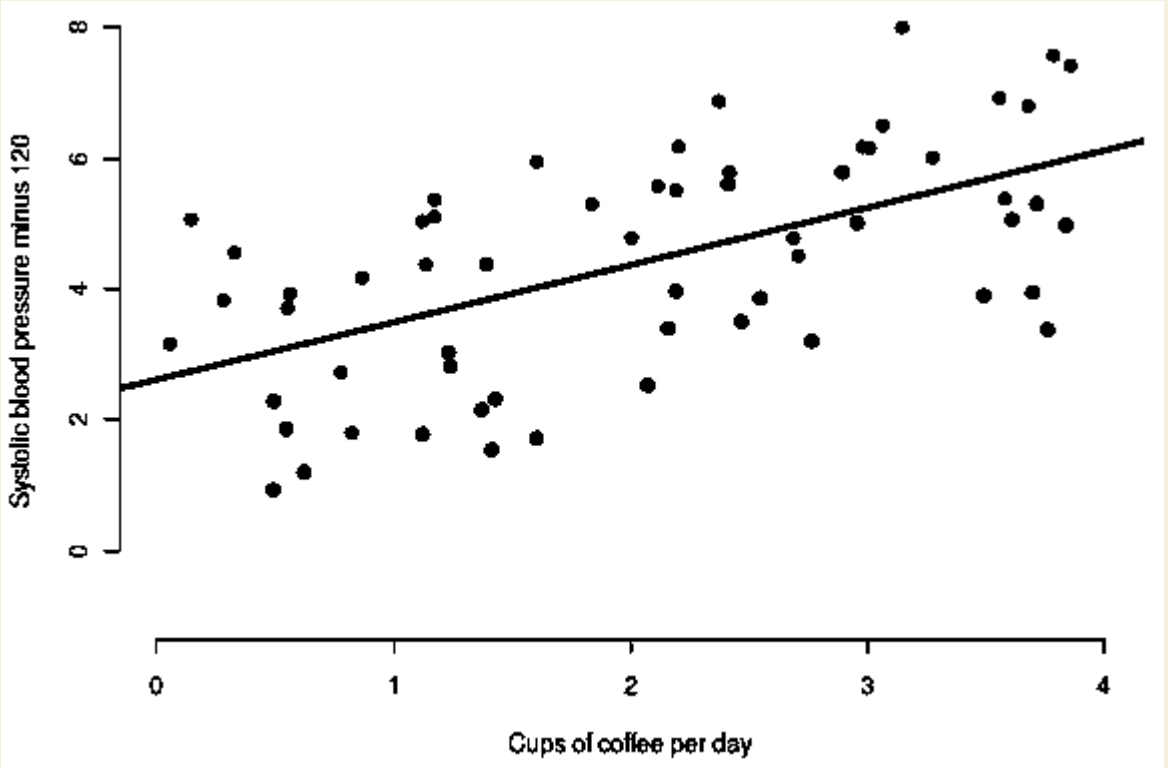
- Evaluer l'effet propre de chaque variable indépendante
- Prendre en compte des facteurs de confusion



Principe

Exemple : étude du lien entre la pression systolique et la consommation de café à partir des données d'un échantillon de personnes issues de la population générale.

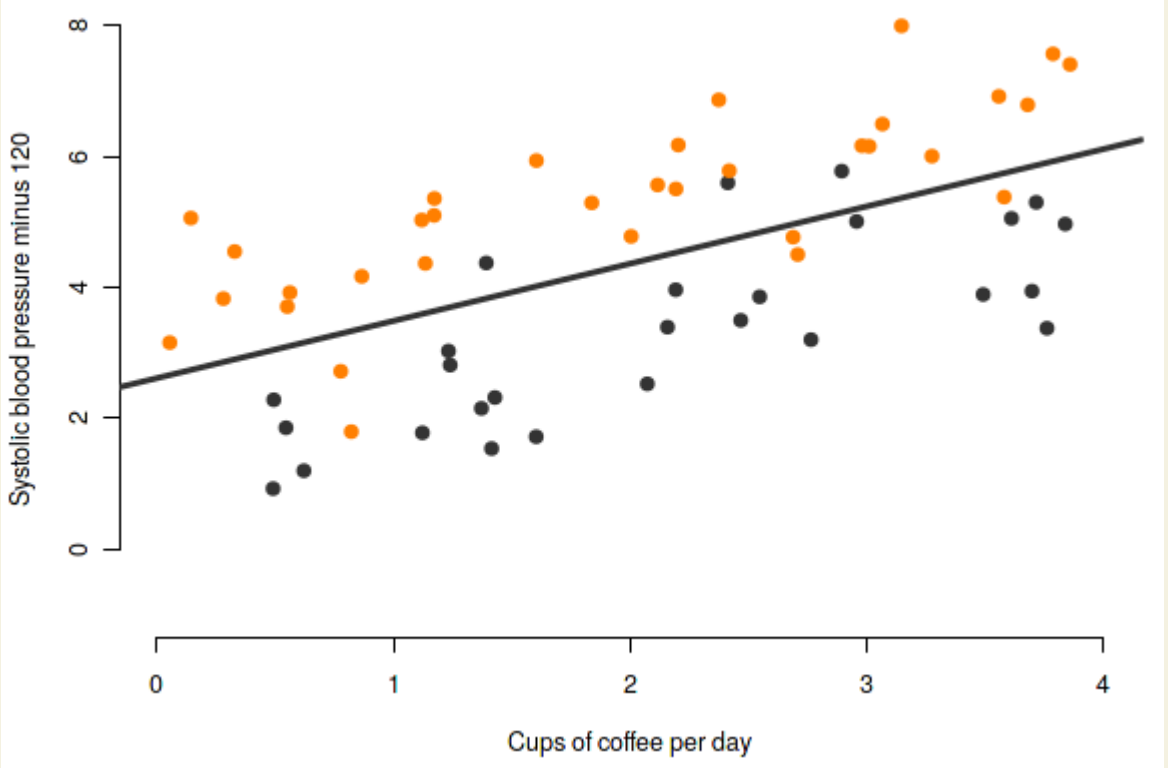
Id	Pression systolique	Nombre de tasses de café par jour	Tabac
1	2,2	1,4	Non fumeur
2	6,0	1,7	Fumeur
…	…	…	



Principe

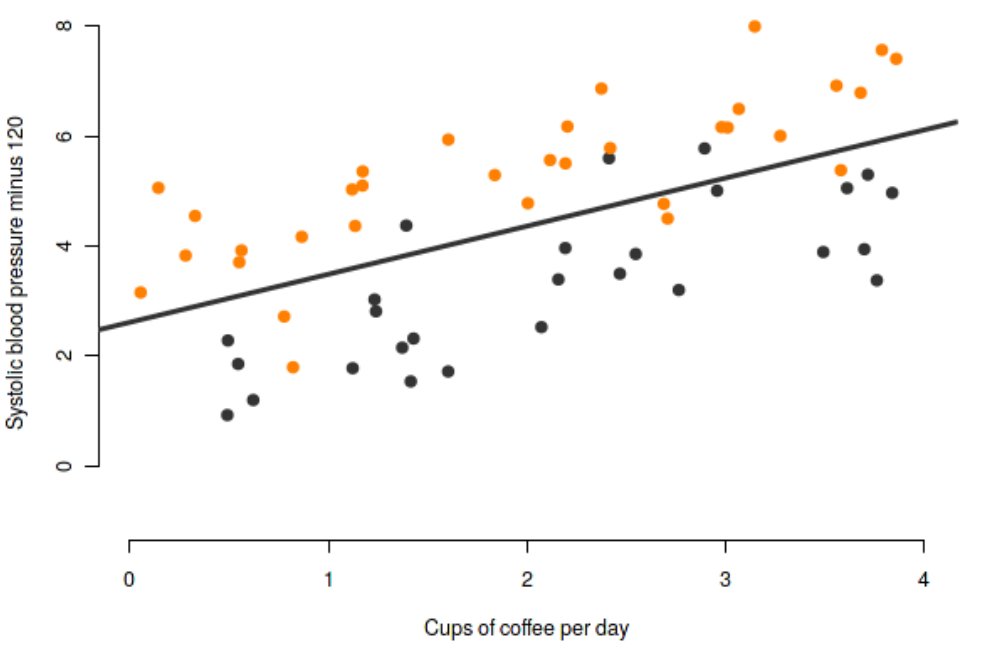
Exemple : étude du lien entre la pression systolique et la consommation de café à partir des données d'un échantillon de personnes issues de la population générale.

Id	Pression systolique	Nombre de tasses de café par jour	Tabac
1	2,2	1,4	Non fumeur
2	6,0	1,7	Fumeur
…	…	…	



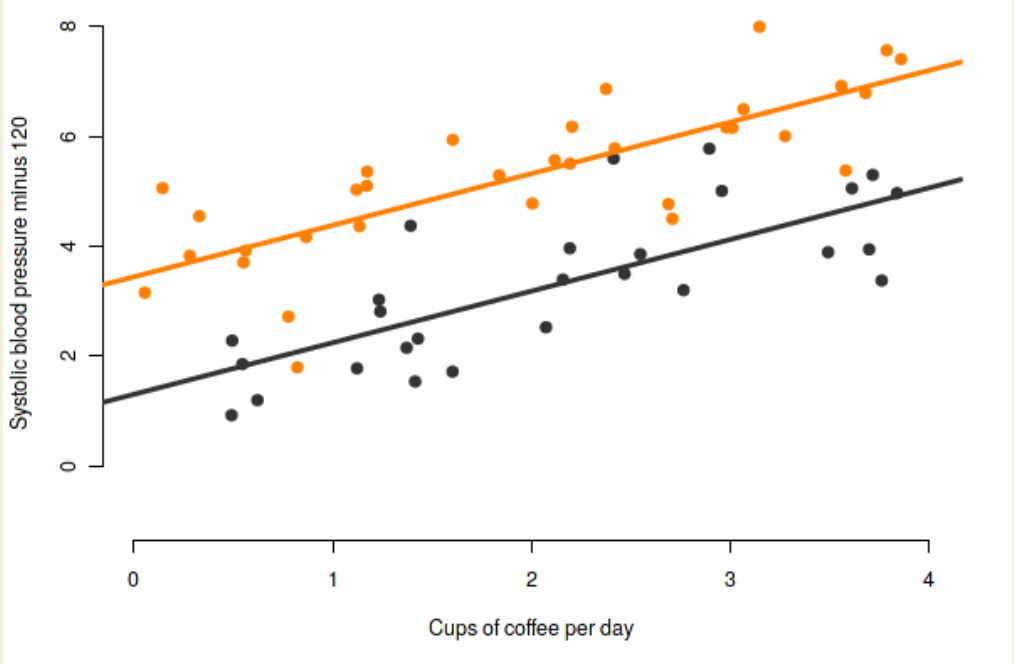
Régression multiple : pourquoi ?

Régression linéaire simple



Variable	Coefficient	p-value
Intercept	2,61	$2x10^{-9}$
Consommation de café	0,87	$7x10^{-7}$

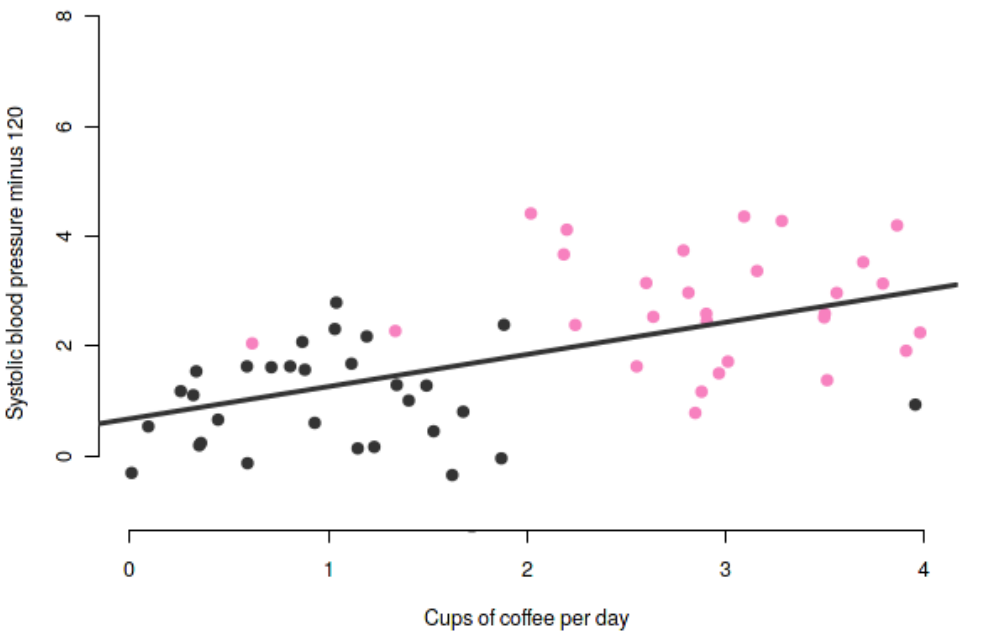
Régression linéaire multiple



Variable	Coefficient	p-value
Intercept	1,30	$1x10^{-5}$
Consommation de café	0,94	$5x10^{-13}$
Statut tabagique	2,14	$7x10^{-13}$

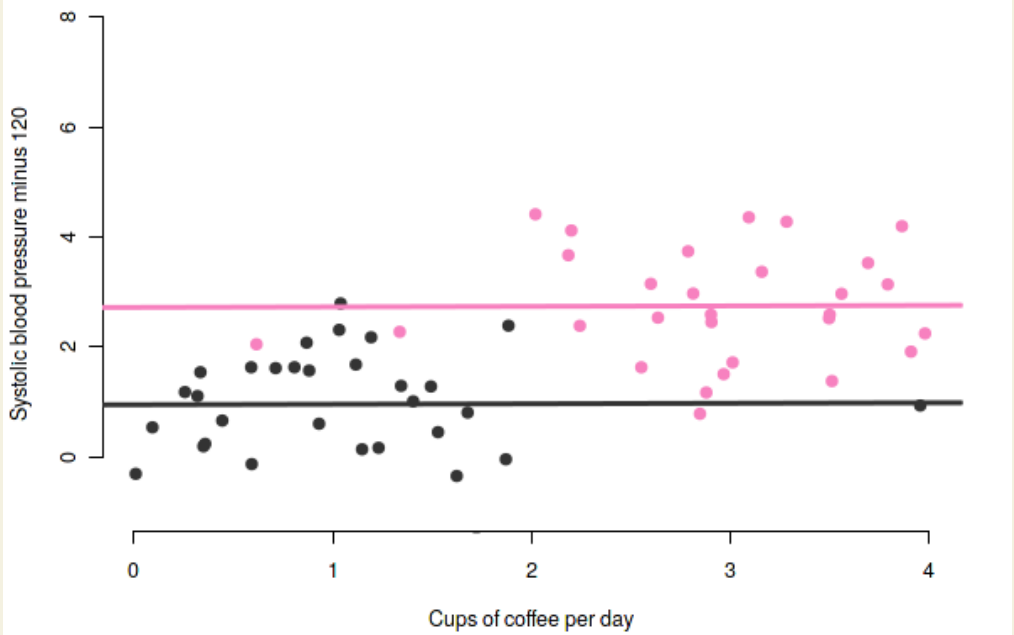
Régression multiple : pourquoi ?

Régression linéaire simple



Variable	Coefficient	p-value
Intercept	0,68	0,019
Consommation de café	0,58	$1x10^{-5}$

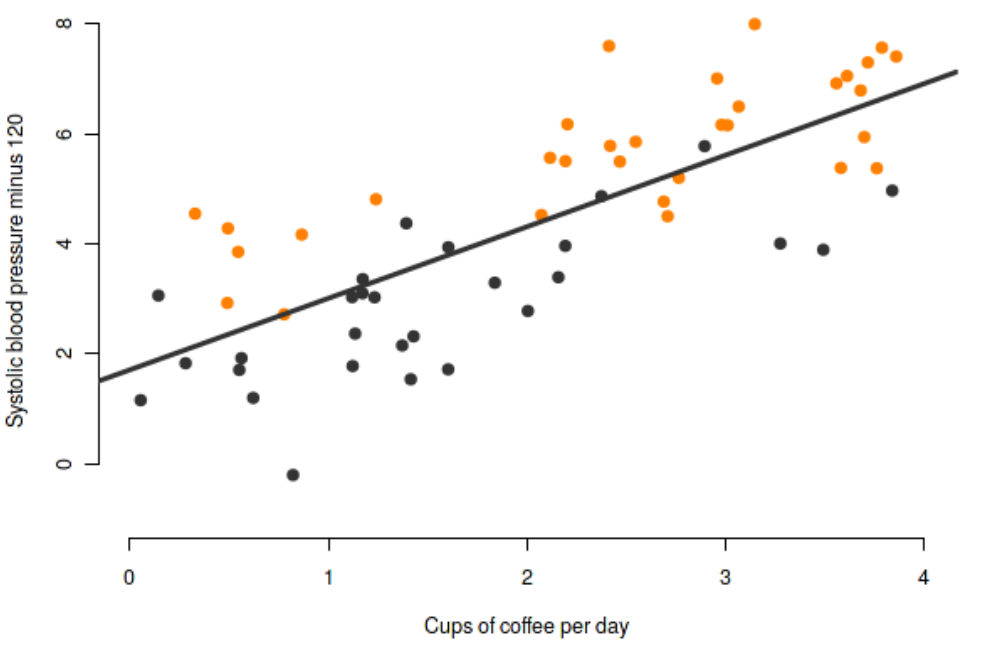
Régression linéaire multiple



Variable	Coefficient	p-value
Intercept	0,95	0,0004
Consommation de café	0,009	0,95
Statut tabagique	1,77	$5x10^{-5}$

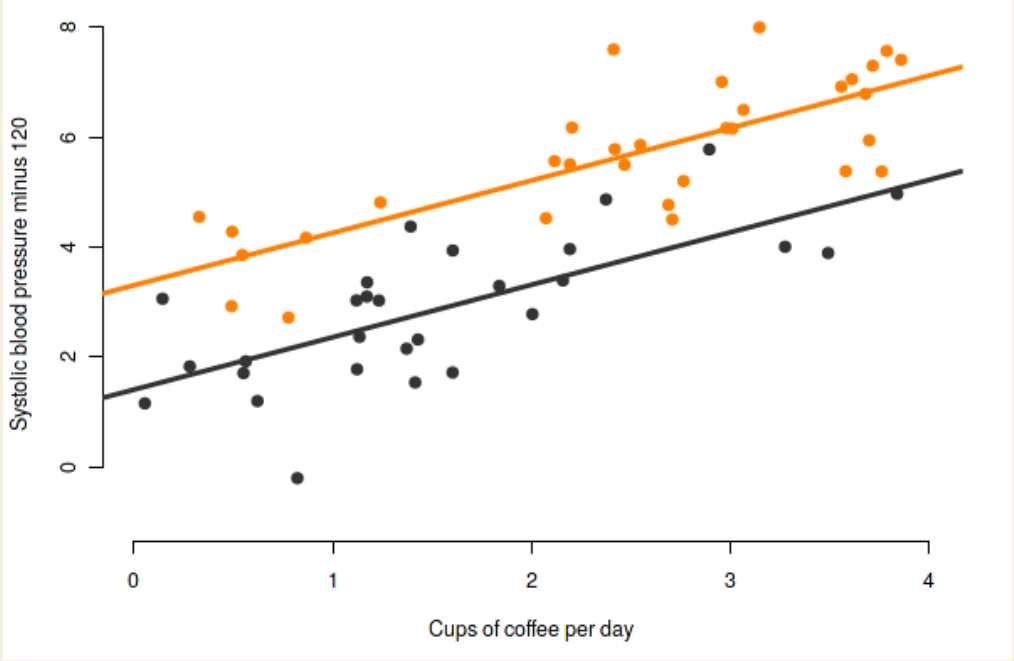
Régression multiple : pourquoi ?

Régression linéaire simple



Variable	Coefficient	p-value
Intercept	1,71	$2x10^{-6}$
Consommation de café	1,30	$5x10^{-13}$

Régression linéaire multiple



Variable	Coefficient	p-value
Intercept	1,40	$2x10^{-7}$
Consommation de café	0,95	$8x10^{-12}$
Statut tabagique	1,90	$6x10^{-10}$

Interprétation des coefficients
Régression multiple

- **α** : moyenne de Y chez les personnes $X_1 = 0, X_2 = 0$, etc.
- Chaque **β** représente l'effet propre de chaque variable **ajusté** sur le niveau des autres variables

$$Y = \alpha + \beta_1 . X_1 + \beta_2 . X_2$$

X = quantitative

β : évolution moyenne de Y lorsque X augmente d'une unité, toutes les autres variables étant égales par ailleurs

X = qualitative binaire

β : écart entre moyenne de Y chez les personnes X=1 et celles X=0, toutes les autres variables étant égales par ailleurs

Test sur les coefficients

- Hypothèses du test :

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

- p-valeur associée à un coefficient < risque 1ère espèce fixé (souvent 5%)

  → Au risque α , la variable X est liée linéairement à Y (après ajustement sur les autres variables explicatives)

- p-valeur associée à un coefficient > risque 1ère espèce fixé

  → Les données (et donc le modèle obtenu) ne permettent pas de montrer que la variable X est liée linéairement à Y (après ajustement sur les autres variables explicatives)

Exercice

Variable	Coefficient
Intercept	2362,50
Statut tabagique pendant la grossesse	-269,26
Age de la mère	7,15
Poids de la mère	4,02

- Ecriture du modèle ?

$$Y = \alpha + \beta_1 . X_1 + \beta_2 . X_2 + \beta_3 . X_3$$

$$\text{POIDS\_BEBE} = 2362{,}50 - 269{,}26 \times \text{TABAC\_GROSSESSE} + 7{,}15 \times \text{AGE\_MERE} + 4{,}02 \times \text{POIDS\_MERE}$$

- Interprétation ?

Exercice

Variable	Coefficient	p-value
Intercept	2362,50	p < 0,0001
Statut tabagique pendant la grossesse	-269,26	0,012
Age de la mère	7,15	0,472
Poids de la mère	4,02	0,021

- Ecriture du modèle ?

$$Y = \alpha + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_3$$

$$\text{POIDS\_BEBE} = 2362{,}50 - 269{,}26 \times \text{TABAC\_GROSSESSE} + 7{,}15 \times \text{AGE\_MERE} + 4{,}02 \times \text{POIDS\_MERE}$$

- Interprétation ?
  - En moyenne, les enfants de mères qui fument pendant leur grossesse pèsent en moyenne 269,26 grammes de moins que les enfants des mères qui ne fument pas pendant leur grossesse.
  - Et ce de façon statistiquement significative (p value = 0,012 < 0,05) au seuil alpha de 5%, après ajustement sur l'âge et le poids de la mère.

Exercice

Variable	Coefficient	p-value
Intercept	2362,50	p < 0,0001
Statut tabagique pendant la grossesse	-269,26	0,012
Age de la mère	7,15	0,472
Poids de la mère	4,02	0,021

- Ecriture du modèle ?

$$Y = \alpha + \beta_1 . X_1 + \beta_2 . X_2 + \beta_3 . X_3$$

$$\text{POIDS\_BEBE} = 2362{,}50 - 269{,}26 \times \text{TABAC\_GROSSESSE} + 7{,}15 \times \text{AGE\_MERE} + 4{,}02 \times \text{POIDS\_MERE}$$

- Interprétation ?
  - Pour une année supplémentaire d'âge de la mère, le poids des enfants à la naissance augmente en moyenne de 7,15 grammes.
  - Mais l'âge de la mère n'est pas significativement associé au poids à la naissance des enfants (p = 0,472 < 0,05)au seuil alpha de 5%, après ajustement sur l'âge et le poids de la mère.

Exercice

Variable	Coefficient	p-value
Intercept	2362,50	p < 0,0001
Statut tabagique pendant la grossesse	-269,26	0,012
Age de la mère	7,15	0,472
Poids de la mère	4,02	0,021

- Ecriture du modèle ?

$$Y = \alpha + \beta_1 . X_1 + \beta_2 . X_2 + \beta_3 . X_3$$

$$\text{POIDS\_BEBE} = 2362{,}50 - 269{,}26 \times \text{TABAC\_GROSSESSE} + 7{,}15 \times \text{AGE\_MERE} + 4{,}02 \times \text{POIDS\_MERE}$$

- Interprétation ?
  - Le poids des enfants à la naissance augmente avec le poids de la mère (4,02 grammes de plus lorsque le poids de la mère augmente d'un kilo).
  - Et ce de façon statistiquement significative (p value = 0,021 < 0,05) au seuil alpha de 5%, après ajustement sur le statut tabagique de la mère pendant la grossesse et son poids.

Hypothèses de la régression linéaire

- Indépendance des variables explicatives
- Linéarité de la relation entre Y et X
- Normalité des résidus
- Homoscédasticité des résidus

} A vérifier à posteriori

Limites

- Conditions d'utilisation souvent peu respectées
- Limitée aux variables dépendantes (Y) quantitatives

  → En recherche médicale, Y est souvent une variable qualitative **binaire**

  Exemples :

  - Décès / Survie
  - Malade ou non
  - Réussite versus échec de traitement
  - Survenue d'événement indésirables ou pas

RÉGRESSION LOGISTIQUE

Principe

- Y = qualitative **binaire** (ex : M^+, M^-)
- X = qualitative ou quantitative

  → Ce que l'on cherche à estimer : la **probabilité de survenue de la maladie** quand la valeur des variables X est connue

$$P(M^+|X_1, X_2, \dots, X_P)$$

  → Fréquence de la maladie doit être exprimée sous forme de **risque**

- Risque : probabilité de survenue de la maladie au cours d'une période donnée

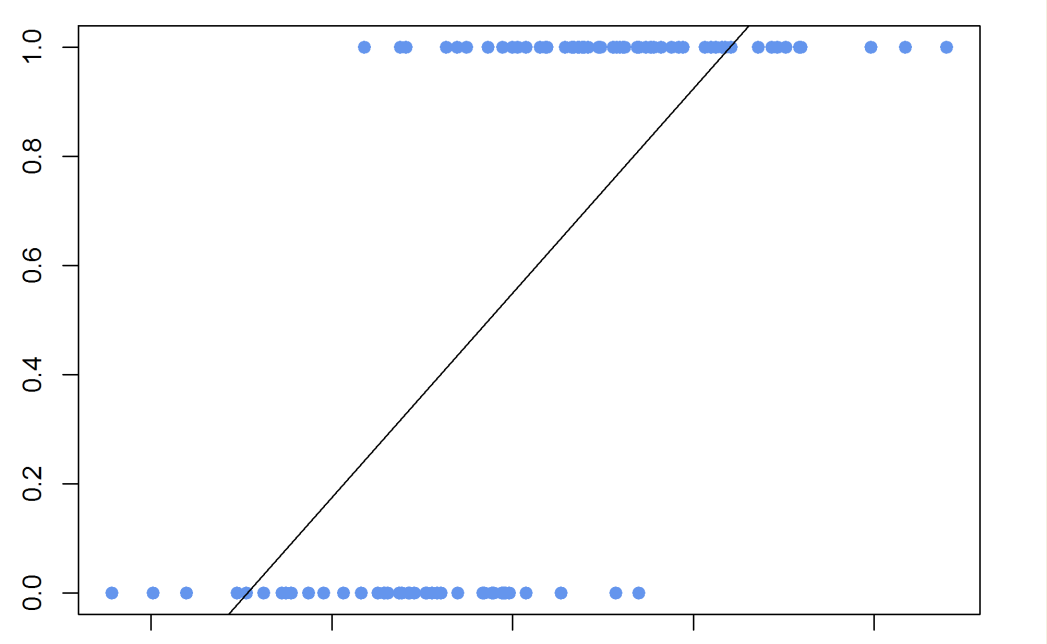
  Exemple : survenue d'un événement médical avec une date seuil (ex : probabilité d'être malade dans les 5 ans)

Fonction logistique

- Quelle fonction ?

$$P(M^+|X) = f(X)$$

→ Si on utilise la régression linéaire :



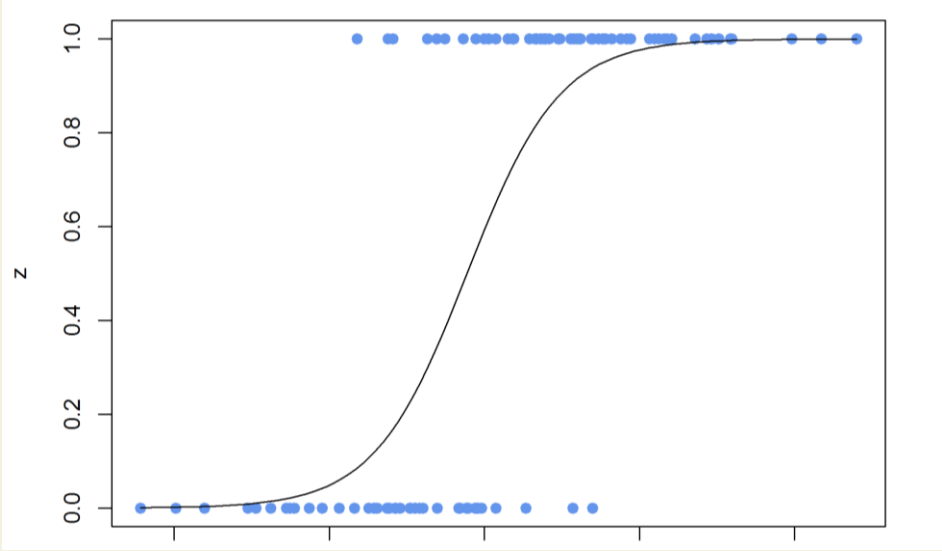
Fonction logistique

- Quelle fonction ?

$$P(M^+|X) = f(X)$$

→ Fonction logistique :

$$f(X) = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$



Odds ratio

- Pourquoi la fonction logistique ?

  → Entre autres, permet de calculer des odds ratio (OR)

- Rappel : odds ratio dans une enquête cas-témoin

	M+	M-
Exposé	a	b
Non exposé	c	d

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Dans une régression logistique :

$$e^{\beta} = OR$$

Interprétation des coefficients

- Comme pour une régression linéaire, on présente :
  - les valeurs estimées des **coefficients β** et de l'**intercept α**
  - ces coefficients étant des estimations réalisées sur un échantillon, on présente l'**écart-type** (ou directement leur **intervalle de confiance**)

Interprétation des coefficients

- α : valeur du logit de la probabilité de survenue de l'événement étudié lorsque les variables explicatives valent 0

  → Comme pour la régression linéaire, n'a pas toujours de sens (ex : si X = âge)

  → Permet de trouver la proportion observée de personnes présentant l'événement étudié pour lesquelles X = 0

$$P(M^+|X=0) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} = \frac{e^{\alpha}}{1+e^{\alpha}}$$

Interprétation des coefficients
Régression simple

Cas 1 : Variable explicative qualitative binaire

→ β : évolution du logit de la probabilité de survenue de l'événement étudié lorsque X passe de 0 à 1

- β positif : être exposé augmente la probabilité de survenue de M par rapport au fait de ne pas l'être
- β négatif : être exposé diminue la probabilité de survenue de M par rapport au fait de ne pas l'être

Exemple : lien entre le risque de survenue du cancer du poumon (Y) et le fait de fumer (X) avec fumeur (X=1) / non-fumeur (X=0).

$$Logit\ P(Y = 1|X) = \alpha + 1{,}2 \times X$$

«Le logit de la probabilité de la survenue d'un cancer est plus élevé de 1,2 chez un fumeur par rapport à un non fumeur… »

⇨ Pas très intuitif en termes d'interprétation

Interprétation des coefficients
Régression simple

→ On préfère donc souvent présenter l'OR correspondant :

$$\mathrm{OR} = e^{\beta}$$

$$OR = e^{1,2} = 3,3$$

⇨ Un fumeur a plus de risque de développer un cancer du poumon qu'un non-fumeur

- Rappels - Interprétation d'un OR :
  - OR > 1 : l'exposition est un facteur de risque de la maladie
  - OR = 1 : l'exposition et la maladie ne sont pas associés
  - OR < 1 : l'exposition est un facteur protecteur de la maladie

Interprétation des coefficients
Régression simple

- Remarque : dans le cas d'une variable explicative qualitative binaire, une régression logistique simple revient au même qu'un test du Chi-deux

  Exemple : poids à la naissance des bébés (Y) en fonction du statut tabagique de la mère pendant la grossesse (X).

	Faible poids à la naissance	Poids « normal » à la naissance
Tabac pendant grossesse	30	44
Pas de tabac pendant grossesse	29	86

  - Résultat du test du Chi-deux : p-value = 0,03
  - Test sur le coefficient de la régression logistique simple : p-value = 0,03

Interprétation des coefficients
Régression simple

Cas 2 : Variable explicative qualitative catégorielle à k modalités

→ Comme pour la régression linéaire ⟶ k variables binaires

→ Un coefficient par variable ; l'interprétation se fait aussi par rapport à la classe de référence

Exemple : fumeur fréquent / fumeur occasionnel / non-fumeur

	Coefficient estimé
Intercept (α)	-1,33
Fumeur fréquent (X_1)	0,66
Fumeur occasionnel (X_2)	0,24

$$Logit\ P(M^+|X) = \alpha + \beta_1.X_1 + \beta_2.X_2$$
$$= -1{,}33 + 0{,}66.X_1 + 0{,}24.X_2$$

Fumeur fréquent par rapport à non-fumeur : OR = $e^{\beta_1} = e^{0,66} = 1,9$

Fumeur occasionnel par rapport à non-fumeur : OR = $e^{\beta_2} = e^{0,24} = 1,3$

Interprétation des coefficients
Régression multiple

Cas 3 : Variable explicative quantitative

→ β : évolution du logit de la probabilité de survenue de l'événement étudié lorsque X augmente d'une unité

- β positif : lorsque X augmente, la probabilité de survenue de M augmente
- β négatif : lorsque X diminue, la probabilité de survenue de M diminue

Exemple : variable X correspondant à l'âge exprimé en années

β = 0,38, soit OR = $e^{\beta} = e^{0,38} = 1,5 > 1$

Si l'âge augmente d'un an, le risque de survenue de la maladie augmente.

→ Limite : suppose que l'OR est le même pour une augmentation de 79 à 80 ans, que pour une augmentation de 25 à 26 ans

- Souvent peu plausible
- Les variables sous forme quantitative sont donc à introduire avec précaution dans un modèle logistique

Interprétation des coefficients
Régression multiple

→ β : évolution du logit de la probabilité de survenue de l'événement étudié lorsque X augmente d'une unité (= logarithme de l'OR correspondant), toutes choses égales par ailleurs

Exemple : étude de la probabilité de survenue du cancer du poumon en fonction de l'âge (X_1), du tabagisme (X_2)

$$\alpha = -1{,}44;\ \beta_1 = 0{,}52; \beta_2 = 0{,}98$$

e^{β} = OR de la variable X correspondante ajusté aux autres variables du modèle

- L'interprétation de α n'a pas de sens
- A âge égal, une personne qui fume a un risque de survenue de la maladie plus élevée qu'une personne qui ne fume pas
- Entre deux personnes ayant le même statut tabagique, la plus jeune a un risque de survenue de la maladie moins élevé que celle plus âgée

Test sur les coefficients

- **β: différent(s) de 0 ?**

  → Là aussi besoin de tester si résultats résultent de fluctuations d'échantillonnage avant de conclure au lien entre la probabilité de survenue P et X

- **Test significatif**

  → OR associé à la variable d'intérêt est statistiquement significatif

- **Test non significatif**

  → On ne peut pas conclure qu'il existe un lien entre la variable d'intérêt et l'événement de santé étudié (OR non significatif)

Exemple

- Etude du lien entre le statut tabagique de la mère durant la grossesse (binaire) et le risque de faible poids à la naissance (<2500 grammes) :

Variable	OR [IC95%]	p-value	β [IC95%]	p-value
Intercept	3,92 [0,56 – 30,69]	0,177	1,36	0,177
Statut tabagique pendant la grossesse	1,95 [1,03 – 3,72]	0,039	0,67	0,039
Age de la mère	0,96 [0,90 – 1,02]	0,233	-0,039	0,233
Poids de la mère	0,97 [0,94 – 0,99]	0,047	-0,026	0,047

Partie redondante avec les 2 premières colonnes (donnent la même information que les OR, mais en moins lisible)

Exemple

- Etude du lien entre le statut tabagique de la mère durant la grossesse (binaire) et le risque de faible poids à la naissance (<2500 grammes) :

Variable	OR [IC95%]	p-value
Intercept	3,92 [0,56 – 30,69]	0,177
Statut tabagique pendant la grossesse	1,95 [1,03 – 3,72]	0,039
Age de la mère	0,96 [0,90 – 1,02]	0,233
Poids de la mère	0,97 [0,94 – 0,99]	0,047

  - Interprétation ?

En moyenne, après ajustement sur les autres paramètres, les mères fumant pendant la grossesse ont un risque significativement plus élevé de présenter un bébé avec un faible poids à la naissance que les femmes ne fumant pas pendant leur grossesse :

- OR = 1,95 > 1
- p-value = 0,039 < 0,05 ou IC95% ne contient pas 1

Exemple

- Etude du lien entre le statut tabagique de la mère durant la grossesse (binaire) et le risque de faible poids à la naissance (<2500 grammes) :

Variable	OR [IC95%]	p-value
Intercept	3,92 [0,56 – 30,69]	0,177
Statut tabagique pendant la grossesse	1,95 [1,03 – 3,72]	0,039
Age de la mère	0,96 [0,90 – 1,02]	0,233
Poids de la mère	0,97 [0,94 – 0,99]	0,047

  - Interprétation ?

En moyenne, le poids des bébés à la naissance diminue avec l'âge de la mère.
Mais, après ajustement sur les autres paramètres, l'âge de la mère n'est pas significativement associé au poids des bébés à la naissance :

- OR = 0,96 < 1
- p-value = 0,233 > 0,05 ou IC95% contient 1

Exemple

- Etude du lien entre le statut tabagique de la mère durant la grossesse (binaire) et le risque de faible poids à la naissance (<2500 grammes) :

Variable	OR [IC95%]	p-value
Intercept	3,92 [0,56 – 30,69]	0,177
Statut tabagique pendant la grossesse	1,95 [1,03 – 3,72]	0,039
Age de la mère	0,96 [0,90 – 1,02]	0,233
Poids de la mère	0,97 [0,94 – 0,99]	0,047

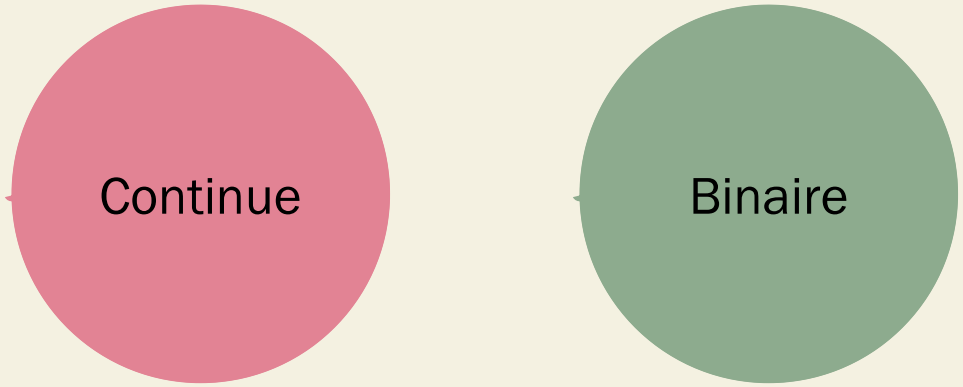
  - Interprétation ?

En moyenne, après ajustement sur les autres paramètres, un poids de la mère plus élevé est associé à un risque significativement plus faible de présenter un bébé avec un faible poids à la naissance :

- OR = 0,97 < 1
- p-value = 0,047 < 0,05 ou IC95% ne contient pas 1

ANALYSE DE SURVIE

Ce que l'on a déjà vu



Continue	Binaire
Poids à la naissance	Survenue d'une maladie
Régression linéaire	Régression logistique

Quelle analyse choisissez-vous ?

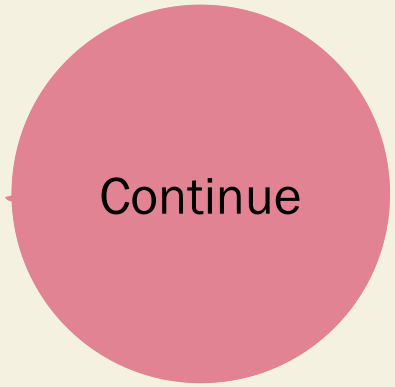
- Etude de l'association entre le délai de rejet d'une greffe de rein et le sexe des personnes.
- Etude de l'association entre la fonction rénale (mesurée par le taux de créatinine) et le sexe des personnes.
- Etude de l'association entre la présence d'un dysfonctionnement rénal ou non suite à un traitement, et le sexe des personnes.

Régression linéaire

Régression logistique

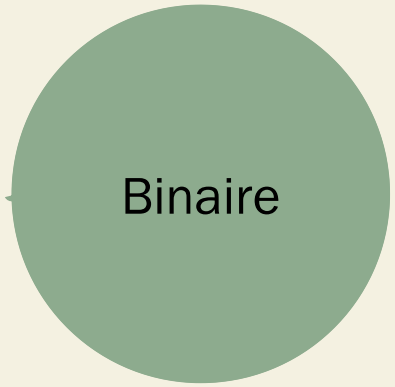
Autre type d'analyse

Ce que l'on va voir



Poids à la naissance

Régression linéaire



Survenue d'une maladie

Régression logistique

Délai jusqu'au décès

Kaplan Meier
Modèle de Cox

Principe

Objectif : étudier le **délai** de survenue d'un événement (binaire)

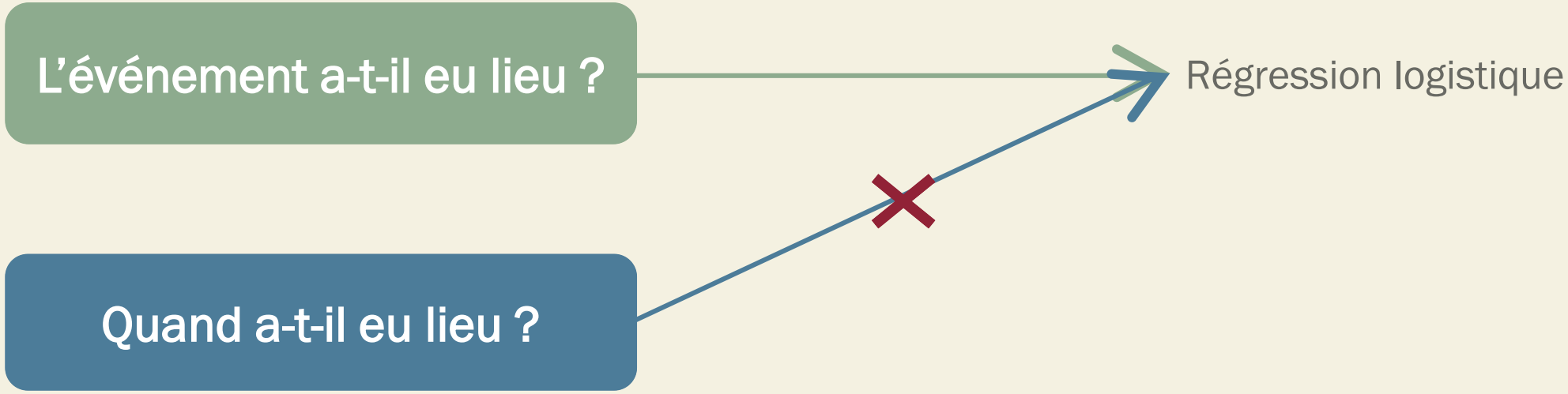


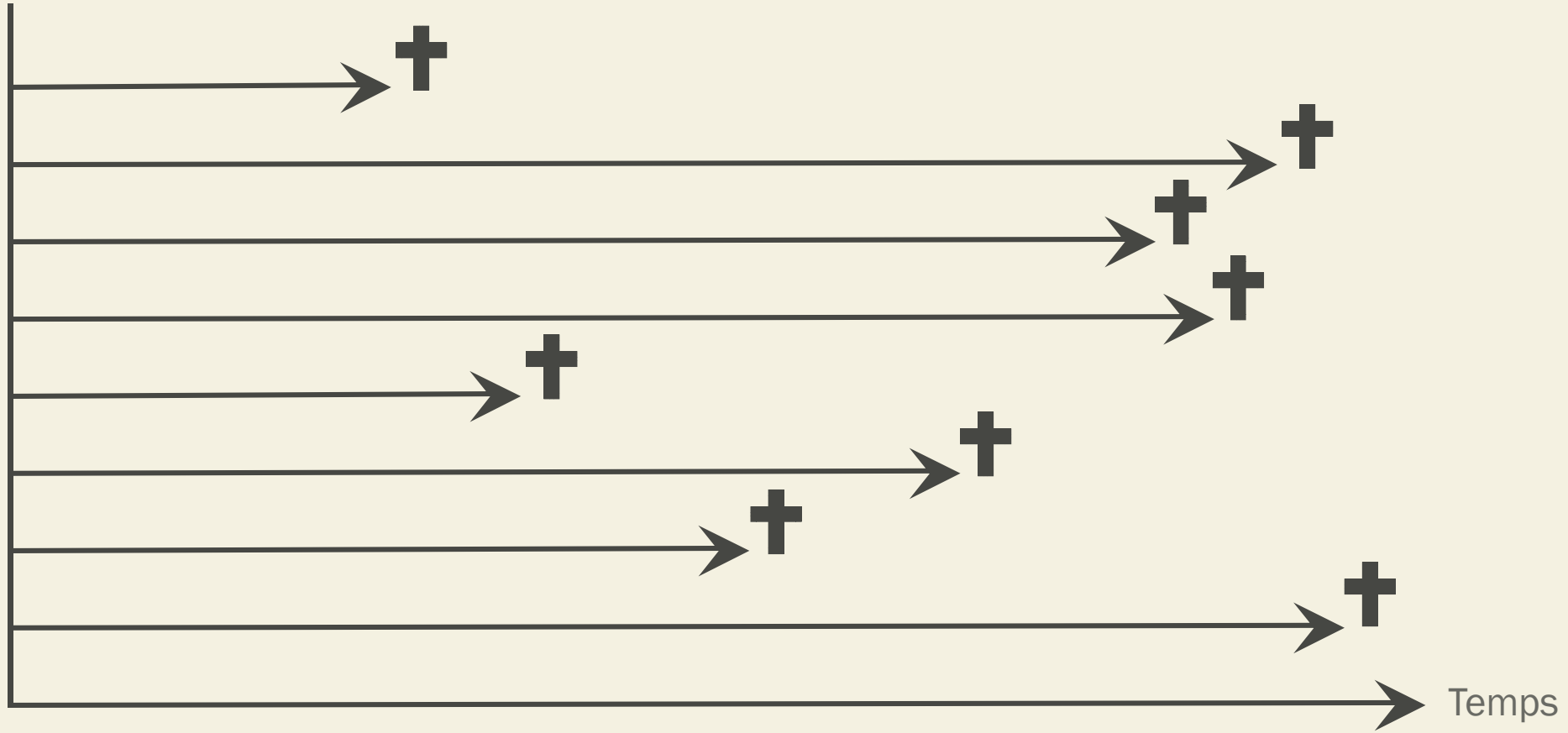
Schéma d’étude

Etude interventionnelle

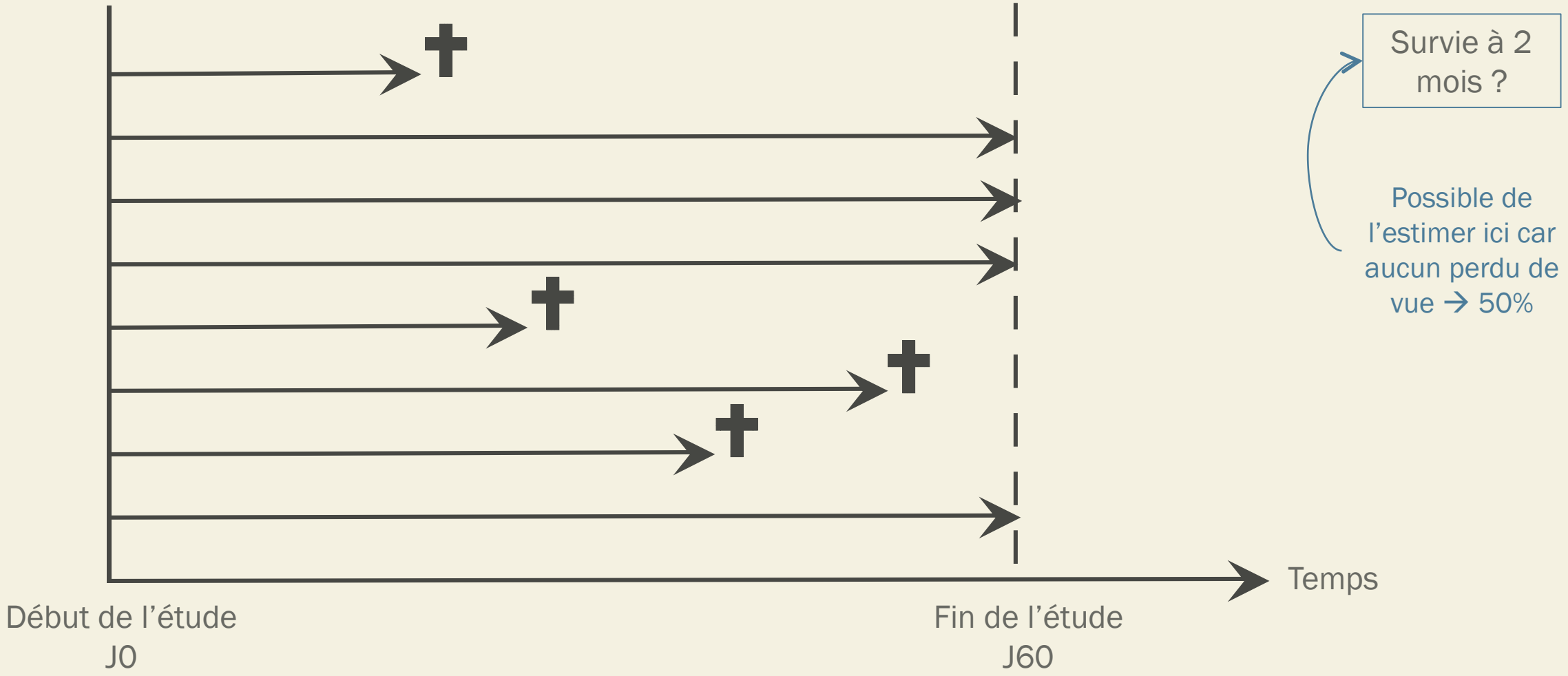
Etude de cohorte

Données censurées

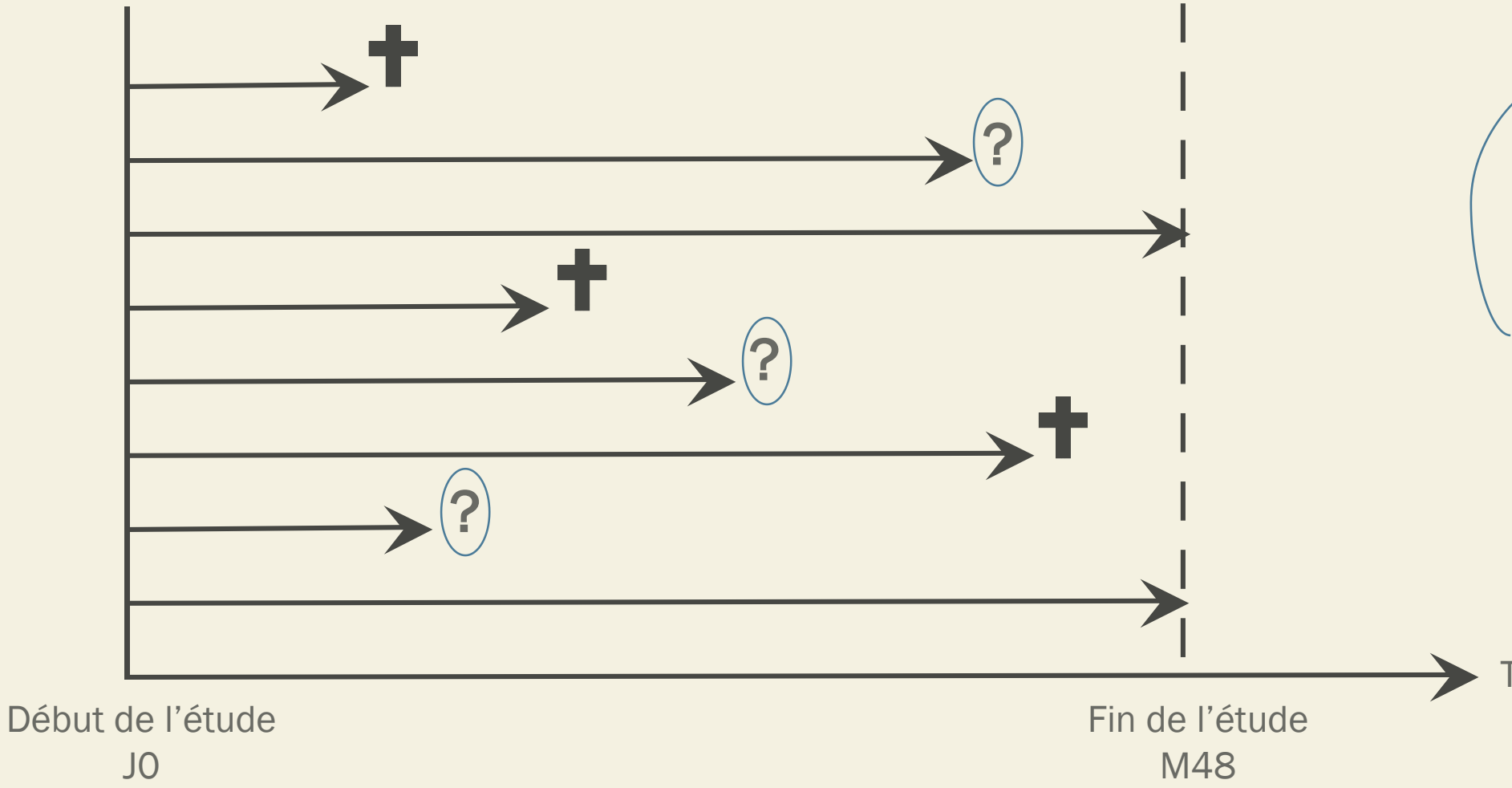
Si l'on suit des personnes toute leur vie, il est possible d'estimer le délai moyen de survenue d'un événement dans la population dont elles sont issue.



Données censurées

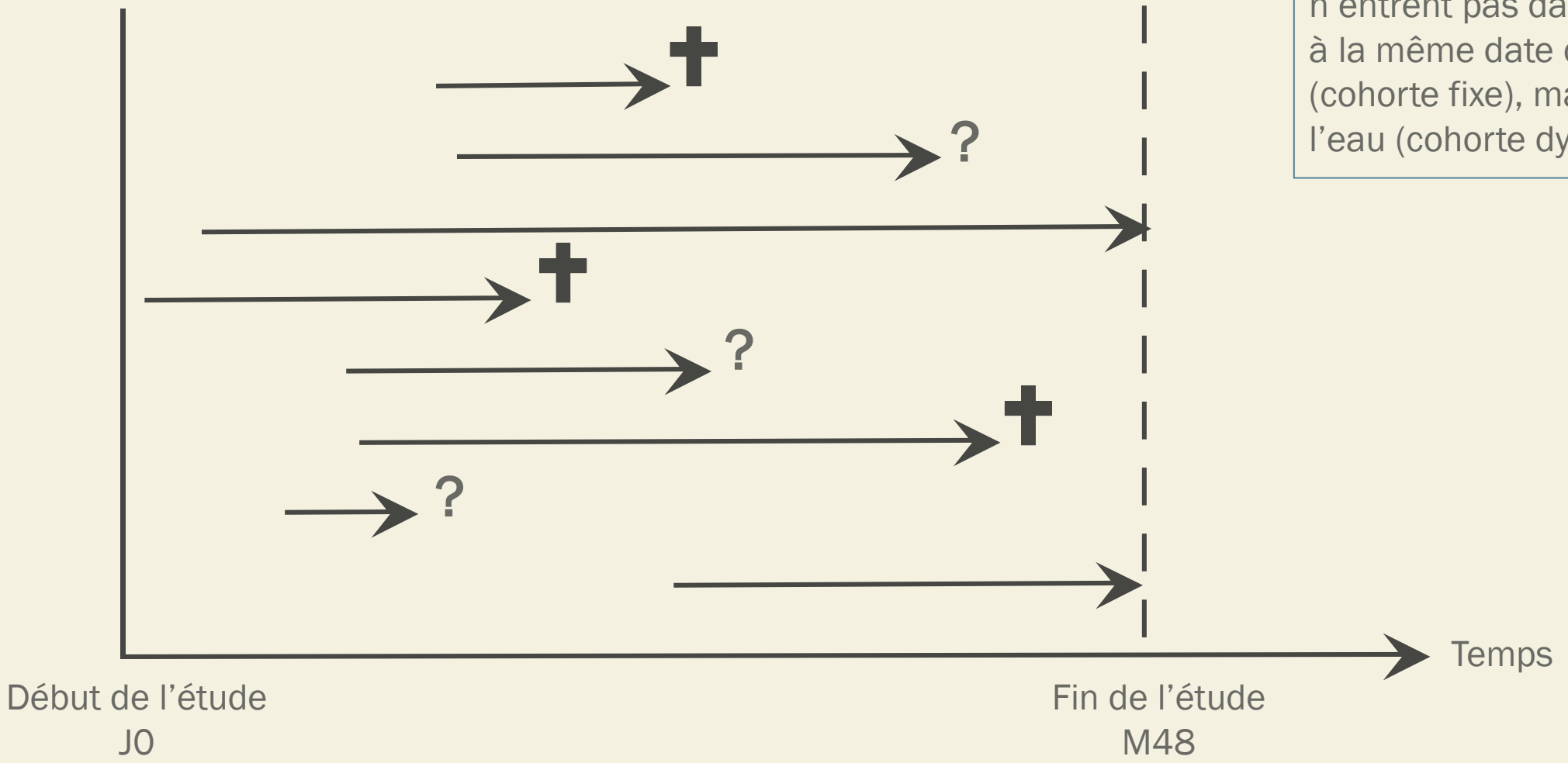


Données censurées

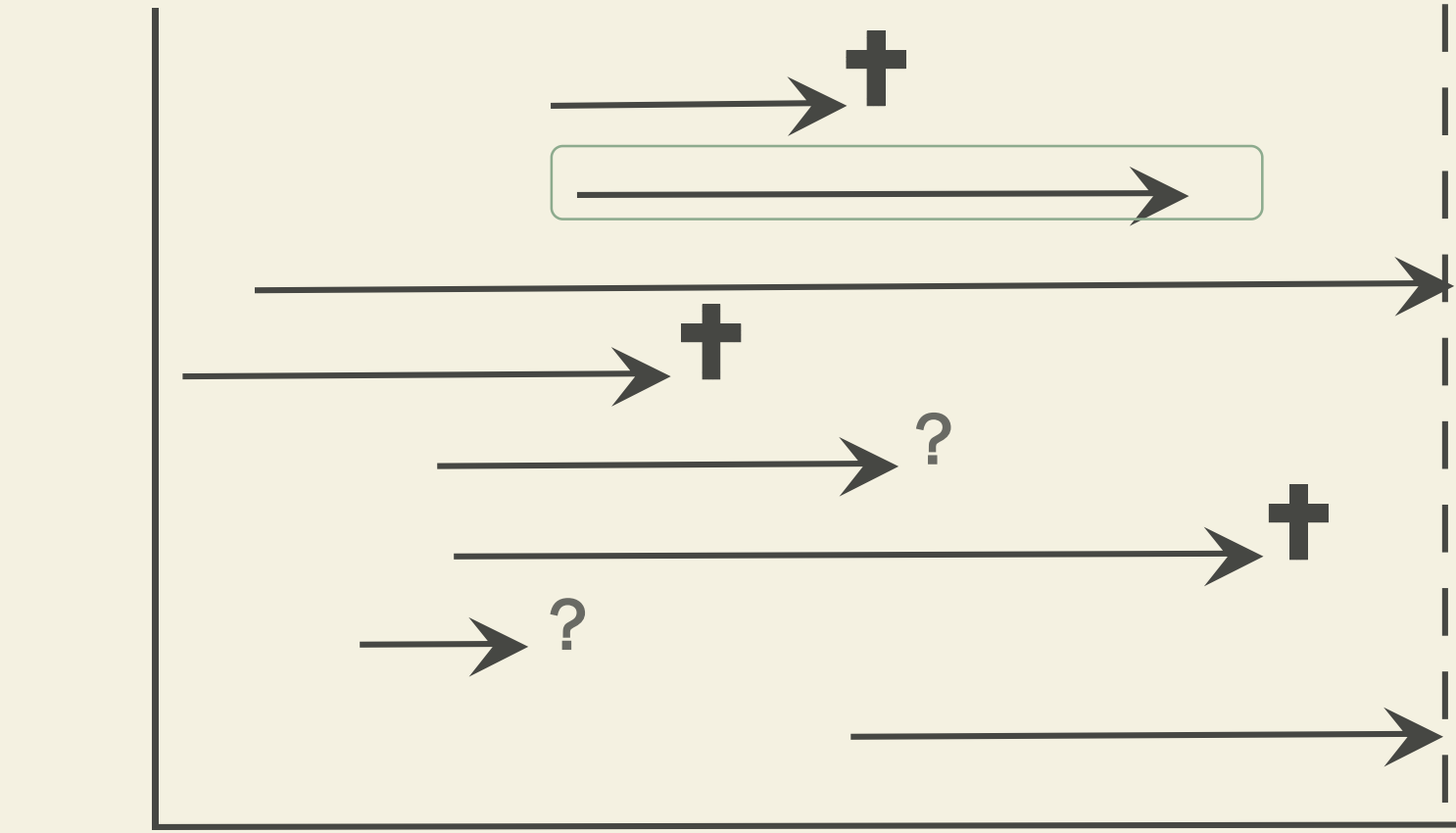


Données censurées

Ici, contrairement aux exemples précédant, les participants de l'étude n'entrent pas dans la cohorte à la même date calendaire (cohorte fixe), mais au fil de l'eau (cohorte dynamique)



Données censurées



3 cas de figure possibles dans les analyses de survie :

- Personnes ayant eu l'événement au cours du suivi
- Perdus de vue
- Exclu-vivants (personnes n'ayant pas eu l'événement au cours du suivi)

Données « censurées » car on ne connait pas le délai d'apparition de l'événement chez ces personnes

Analyse des données censurées

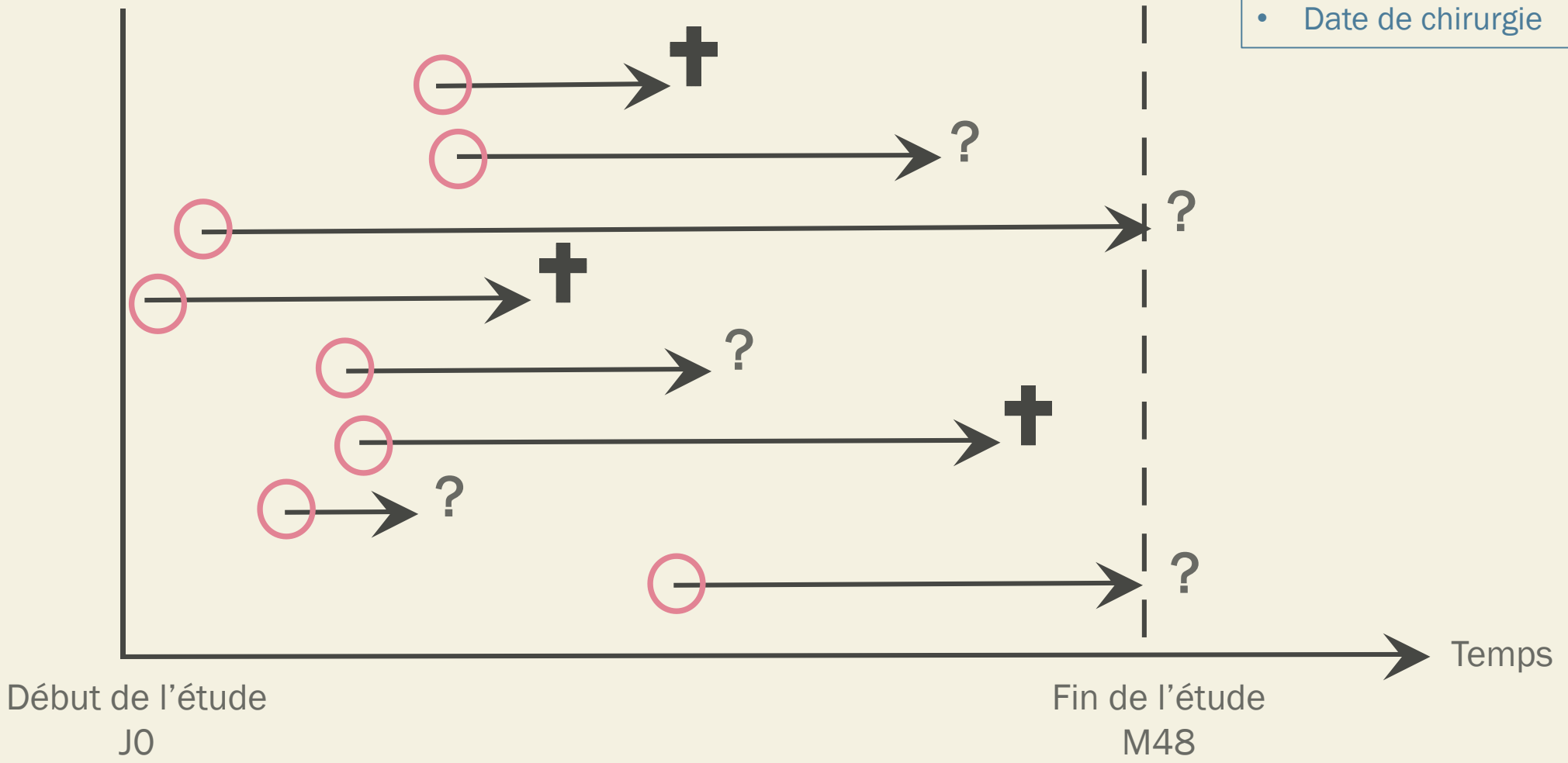
- Régression linéaire ?
  - → Comment inclure les résultats des exclus vivant (= « au moins x jours ») ?
  - → Distribution des délais est souvent asymétrique (invalide l'hypothèse de normalité)
- Ne considérer que les patients ayant l'événement au cours du suivi ?
  - → Biais d'estimation
  - → Perte d'information
- Données censurées nécessitent une analyse particulière
  - → Méthode actuarielle (non abordée ici)
  - → Méthode de Kaplan-Meier

La solution utilisée dans les études de survie est d'estimer des pourcentages de survie tout au long du suivi des patients. Afin de prendre en compte les perdus de vue, les patients seront considérés uniquement sur leur durée de participation à l'étude.

Date d'origine

= Date de début de suivi du sujet. A la même définition pour tous les sujets.
Exemples :

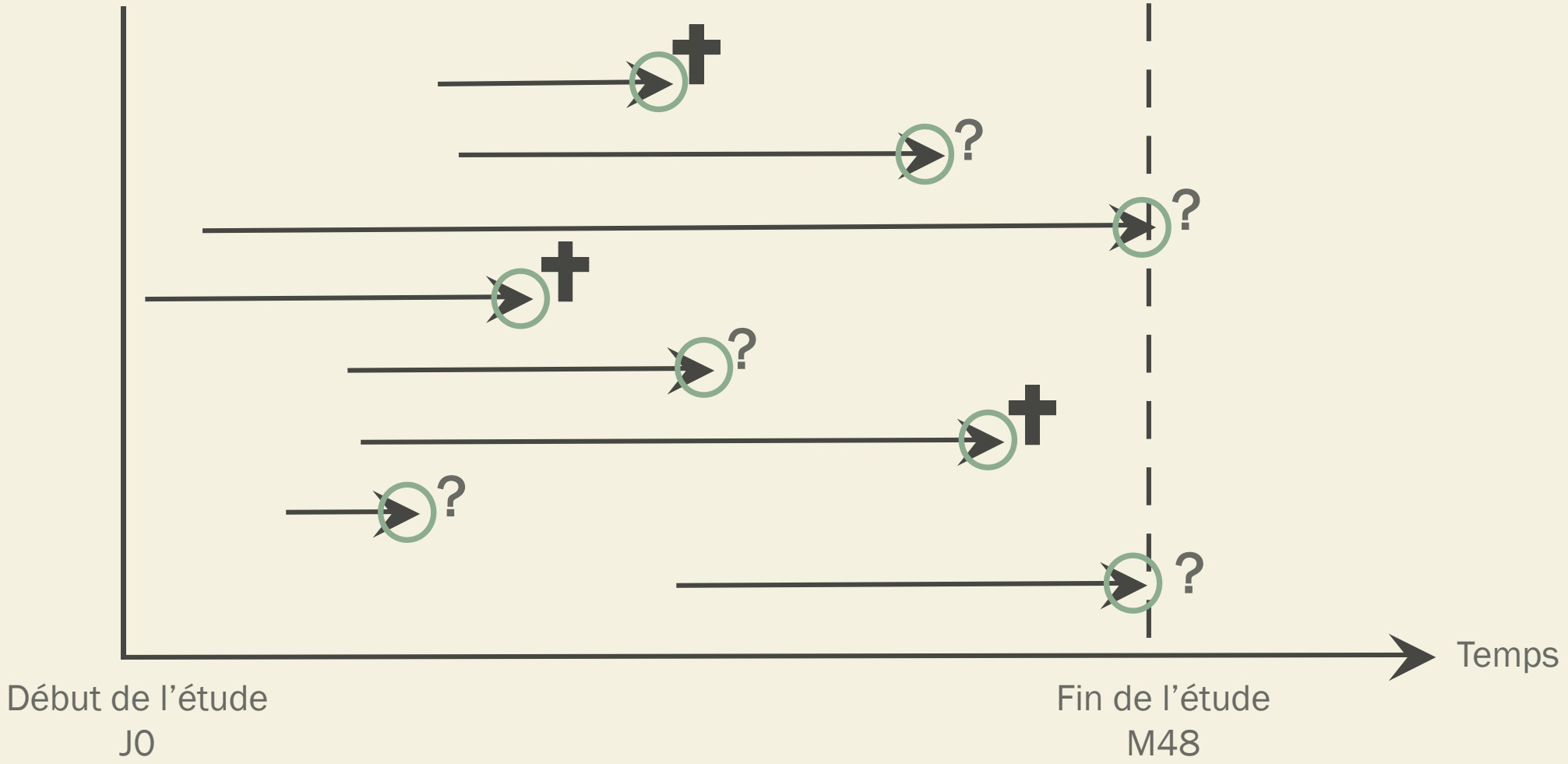
- Date d'inclusion dans un essai,
- Date de diagnostic
- Date de la 1ère métastase
- Date de chirurgie



Date de dernières nouvelles

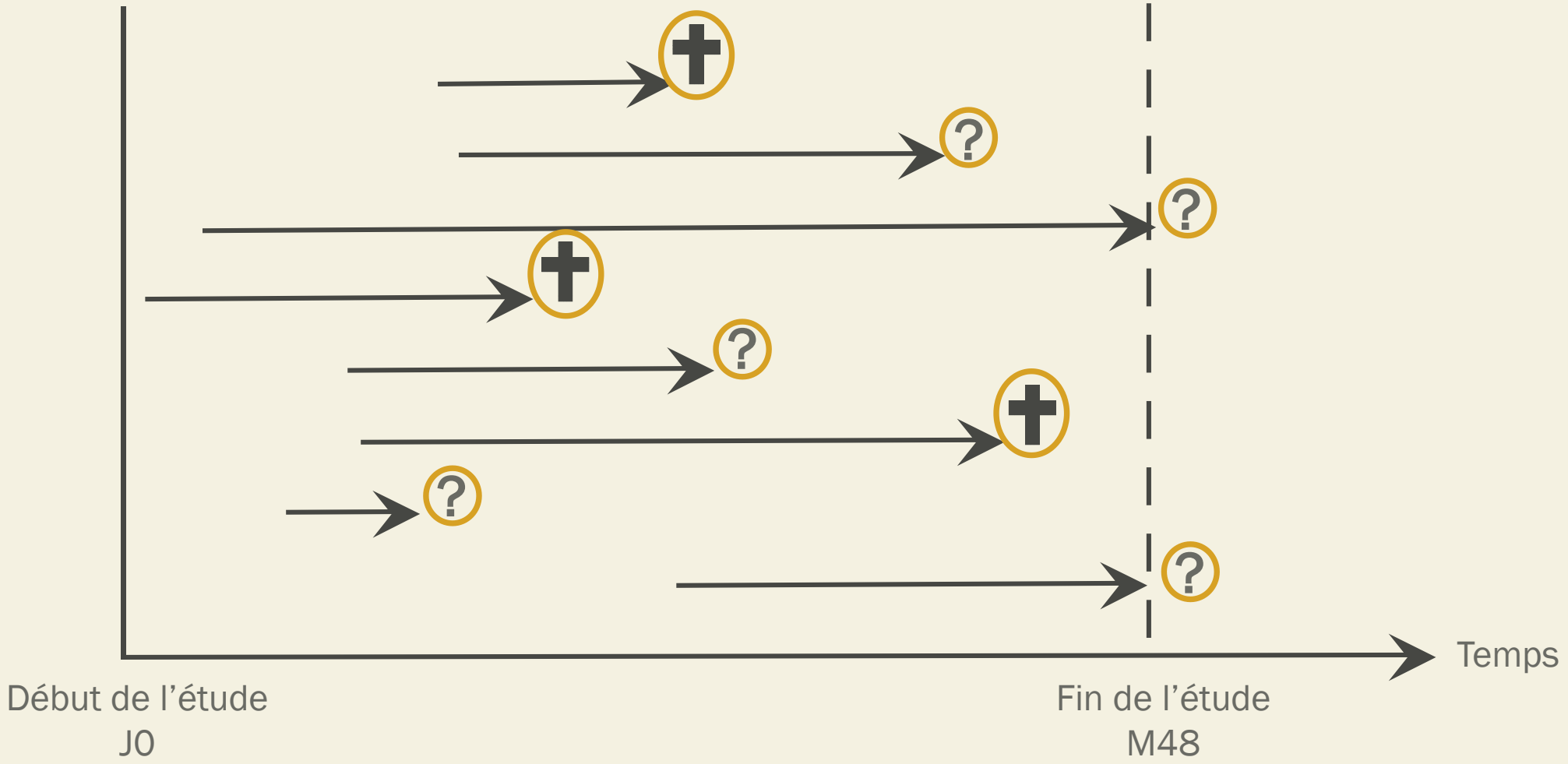
= Date la plus récente où on a observé état du sujet

- Date de l'événement si a eu lieu
- Dernière date où on a observé état du sujet (si l'événement n'a pas eu lieu)



Etat aux dernières nouvelles

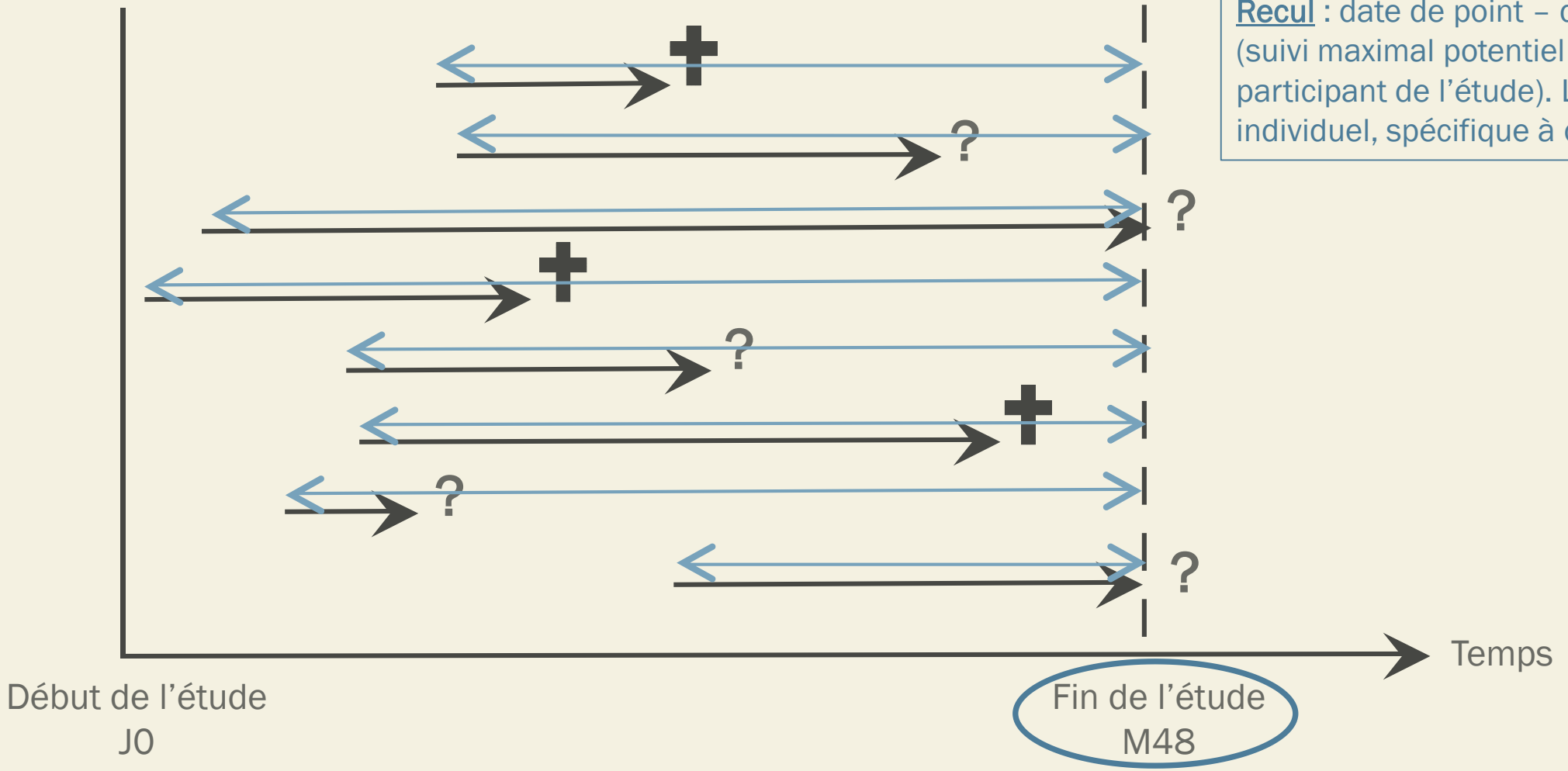
= Présence ou absence de l'événement d'intérêt.
A la même définition pour tous les sujets



Date de point et recul

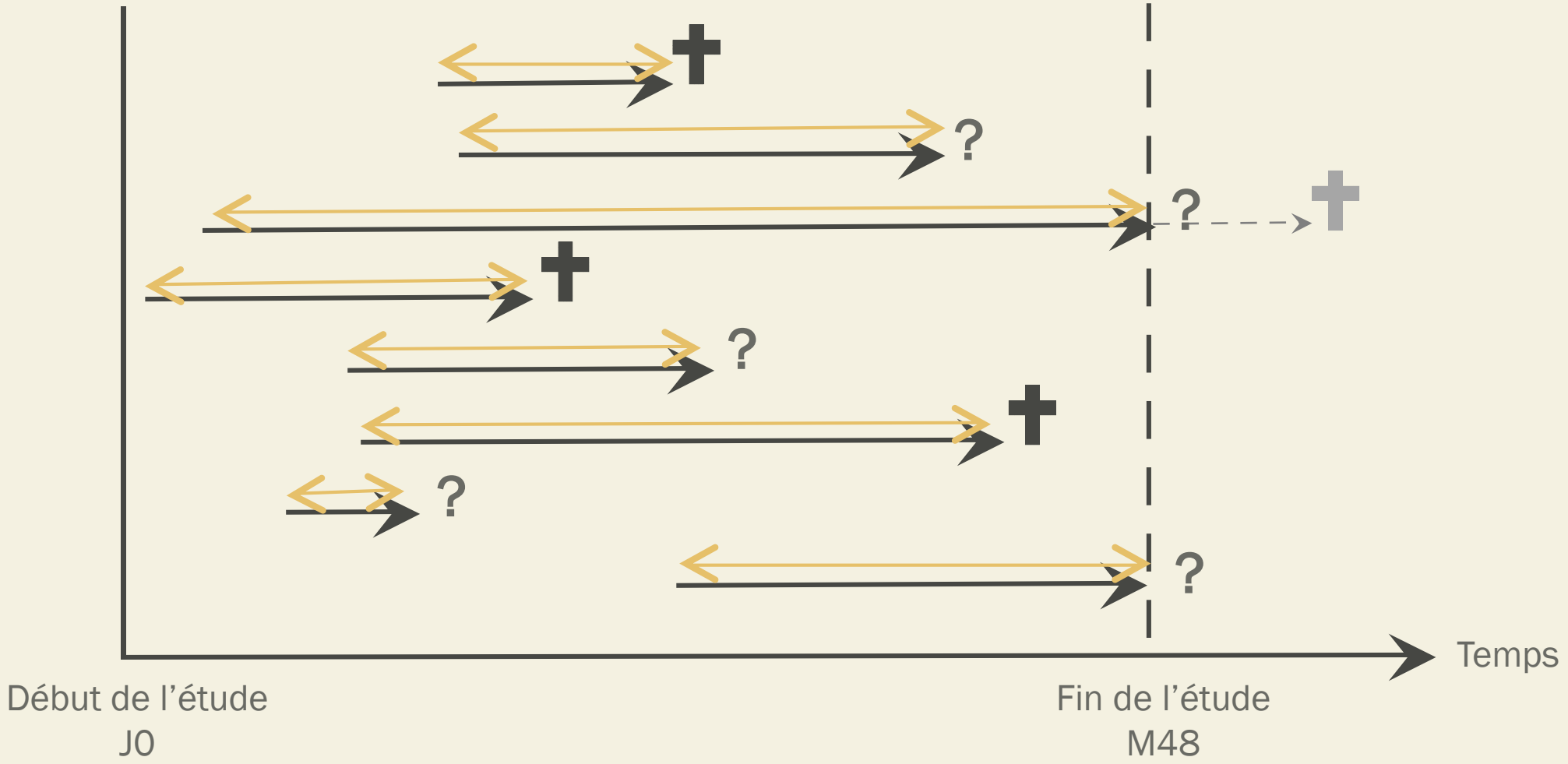
Date de point = date à laquelle on arrête de prendre en compte les événements
- Commune à tous les sujets
- La plus proche possible du moment de l'analyse statistique

Recul : date de point – date d'origine (suivi maximal potentiel de chaque participant de l'étude). Le recul est individuel, spécifique à chaque patient.



= date de dernières nouvelles – date d'origine

Temps de participation



Exercice

Date de Point = 01/03/2020

n° sujet	Date Origine (DO)	Etat à Date de Point (DP)	Date Dernières Nouvelles (DDN)	Etat aux dernières nouvelles	Temps de participation t_i (mois)	Evénement
1	01/03/2019	Vivant	01/08/2020	Vivant		
2	01/01/2020	Vivant	01/10/2020	Décédé		
3	01/04/2019	Perdu de vue (PDV)	01/01/2020	Vivant		
4	01/12/2018	Décédé	01/11/2019	Décédé		

Exercice

Date de Point = 01/03/2020

n° sujet	Date Origine (DO)	Etat à Date de Point (DP)	Date Dernières Nouvelles (DDN)	Etat aux dernières nouvelles	Temps de participation t_i (mois)	Evénement
1	01/03/2019	Vivant	01/08/2020	Vivant	12	Non
2	01/01/2020	Vivant	01/10/2020	Décédé		
3	01/04/2019	Perdu de vue (PDV)	01/01/2020	Vivant		
4	01/12/2018	Décédé	01/11/2019	Décédé		

Exercice

Date de Point = 01/03/2020

n° sujet	Date Origine (DO)	Etat à Date de Point (DP)	Date Dernières Nouvelles (DDN)	Etat aux dernières nouvelles	Temps de participation t_i (mois)	Evénement
1	01/03/2019	Vivant	01/08/2020	Vivant	12	Non
2	01/01/2020	Vivant	01/10/2020	Décédé	2	Non
3	01/04/2019	Perdu de vue (PDV)	01/01/2020	Vivant		
4	01/12/2018	Décédé	01/11/2019	Décédé		

Exercice

Date de Point = 01/03/2020

n° sujet	Date Origine (DO)	Etat à Date de Point (DP)	Date Dernières Nouvelles (DDN)	Etat aux dernières nouvelles	Temps de participation t_i (mois)	Evénement
1	01/03/2019	Vivant	01/08/2020	Vivant	12	Non
2	01/01/2020	Vivant	01/10/2020	Décédé	2	Non
3	01/04/2019	Perdu de vue (PDV)	01/01/2020	Vivant	9	Non
4	01/12/2018	Décédé	01/11/2019	Décédé		

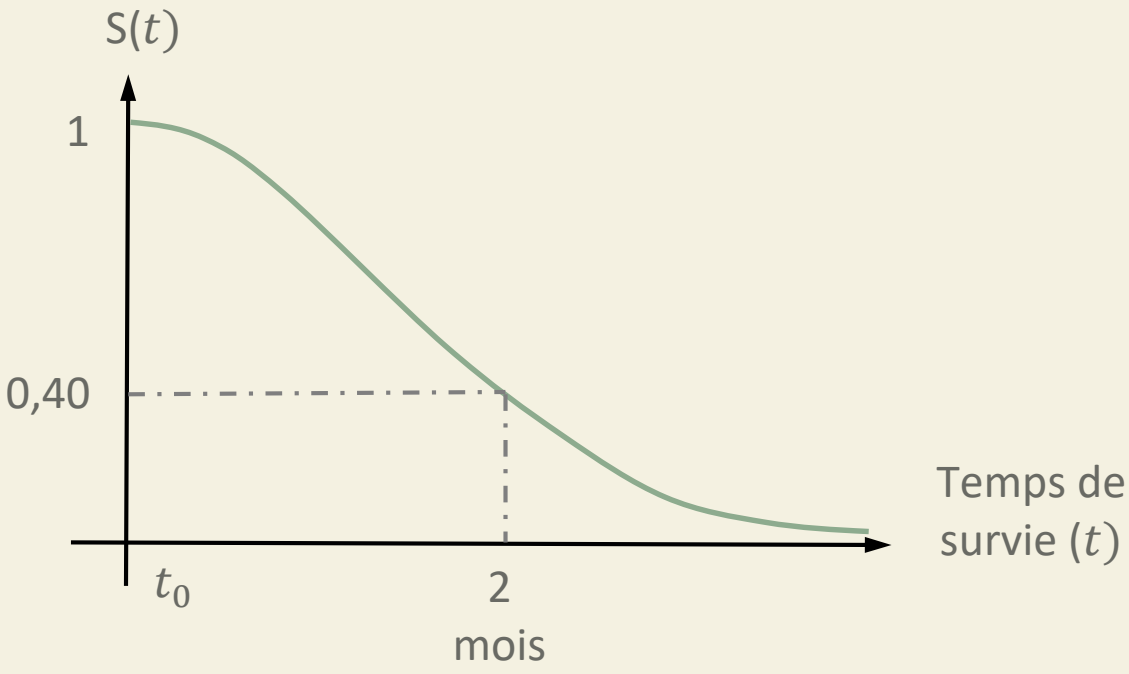
Exercice

Date de Point = 01/03/2020

n° sujet	Date Origine (DO)	Etat à Date de Point (DP)	Date Dernières Nouvelles (DDN)	Etat aux dernières nouvelles	Temps de participation t_i (mois)	Evénement
1	01/03/2019	Vivant	01/08/2020	Vivant	12	Non
2	01/01/2020	Vivant	01/10/2020	Décédé	2	Non
3	01/04/2019	Perdu de vue (PDV)	01/01/2020	Vivant	9	Non
4	01/12/2018	Décédé	01/11/2019	Décédé	11	Oui

Courbe de Kaplan Meier

On cherche à estimer la **fonction de survie.**



$$S(t) = P(T \geq t)$$

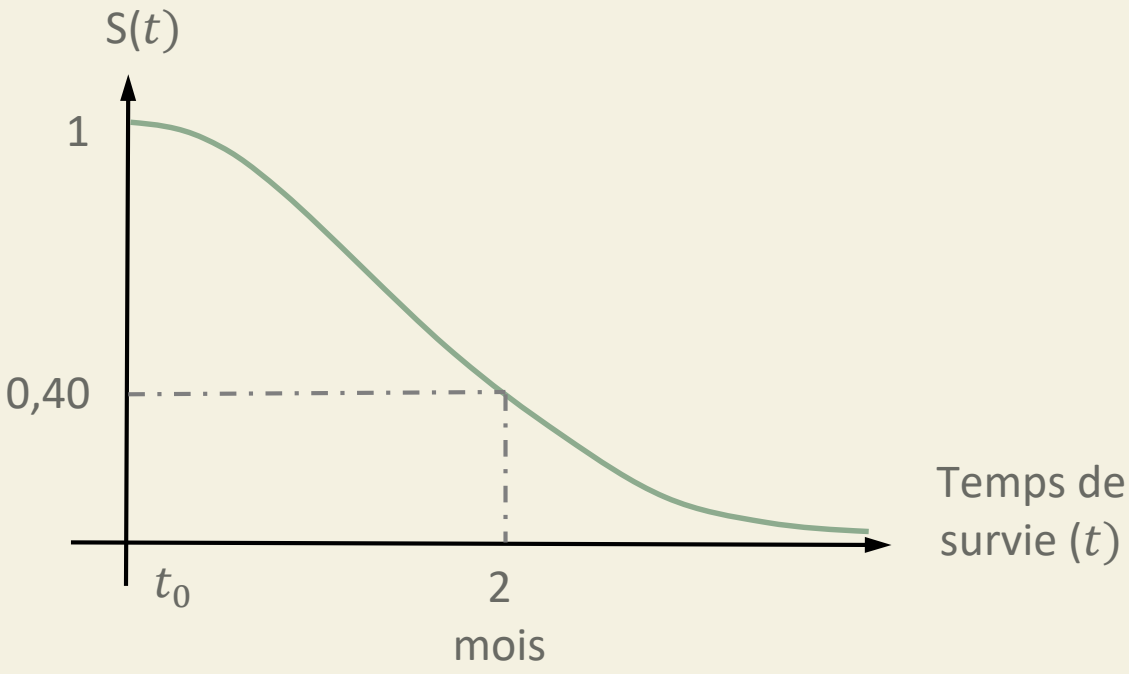
En l'absence de censure :

$$S(t) = \frac{nb\ de\ personnes\ encore\ en\ vie\ après\ t}{nb\ de\ personnes\ à\ risque\ au\ début\ de\ l'étude}$$

→ La probabilité d'être indemne de l'événement considéré 2 mois après t_0 est de 40%

Courbe de Kaplan Meier

On cherche à estimer la **fonction de survie.**



$$S(t) = P(T \geq t)$$

En l'absence de censure :

$$S(t) = \frac{nb\ de\ personnes\ encore\ en\ vie\ après\ t}{nb\ de\ personnes\ à\ risque\ au\ début\ de\ l'étude}$$

En présence de censure :

→ Nombre de personnes à risque diminue au cours du temps

→ Il faut procéder par période de temps pour estimer S(t)

Estimation
Méthode de Kaplan-Meier

Démarche :

1. Classer les sujets par temps de participation individuelle t_i croissant
2. A chaque instants où un événement est observé
    - Identifier les sujets à risque
    - Compter les événements pour l'intervalle de temps considéré
    - Estimer S(t)

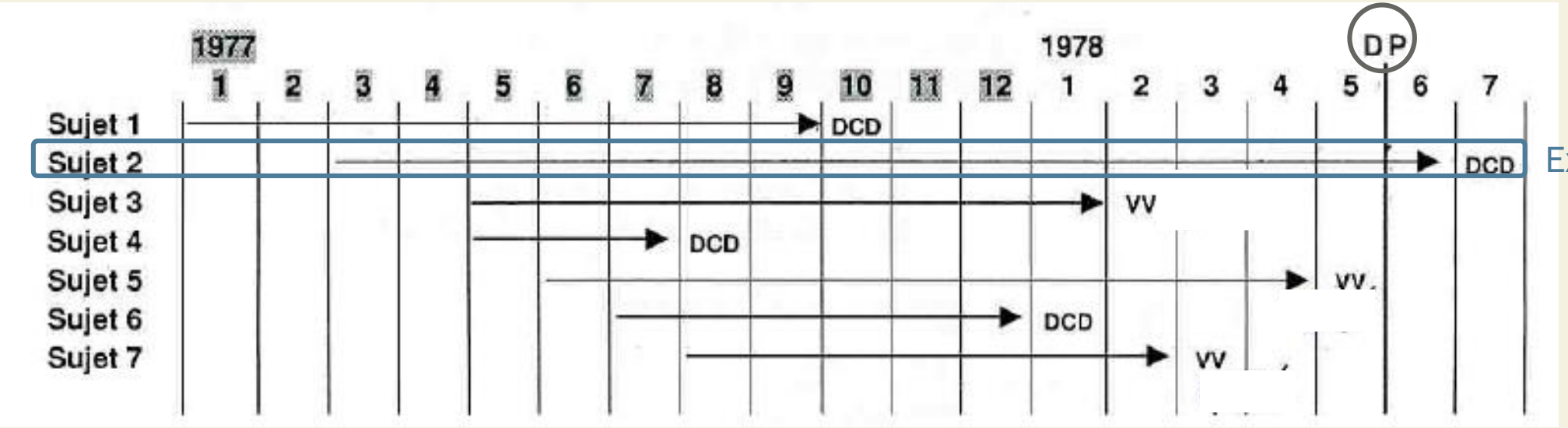
→ *La fonction de survie S(t) est réévaluée à chaque événement*

→ *S(t) est considérée comme constante entre deux événements*

Estimation
Méthode de Kaplan-Meier

<u>Exemple</u> : 7 sujets dans une étude thérapeutique

→ Date de Point (DP) = 01/06/1978



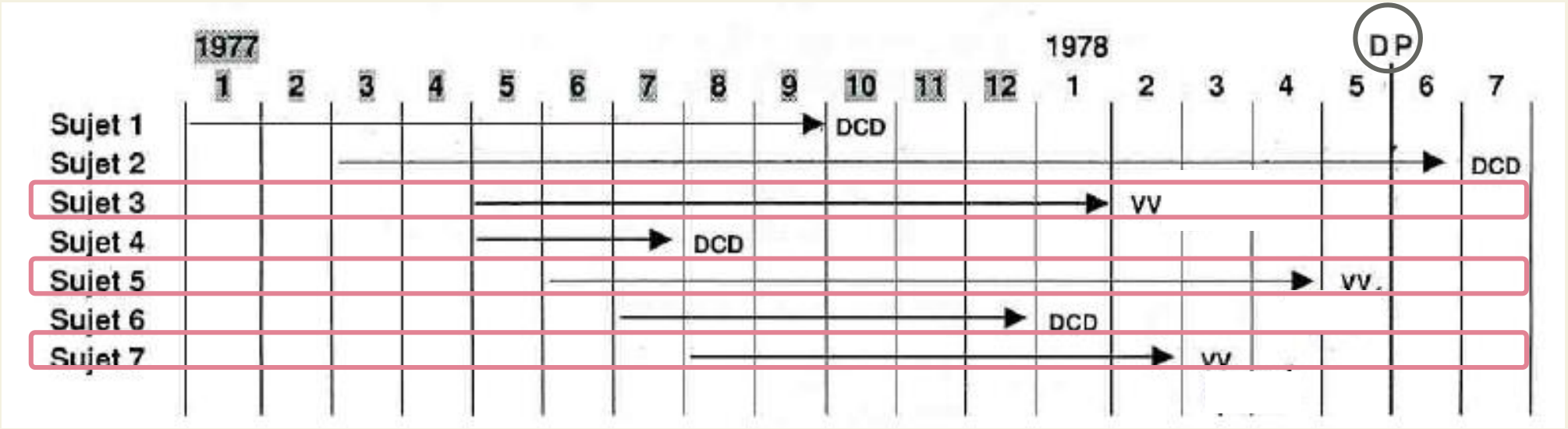
Exclu-vivant

DCD = décédé
VV = vivant

Estimation
Méthode de Kaplan-Meier

<u>Exemple</u> : 7 sujets dans une étude thérapeutique

→ Date de Point (DP) = 01/06/1978



PDV

DCD = décédé
VV = vivant

Estimation
Méthode de Kaplan-Meier

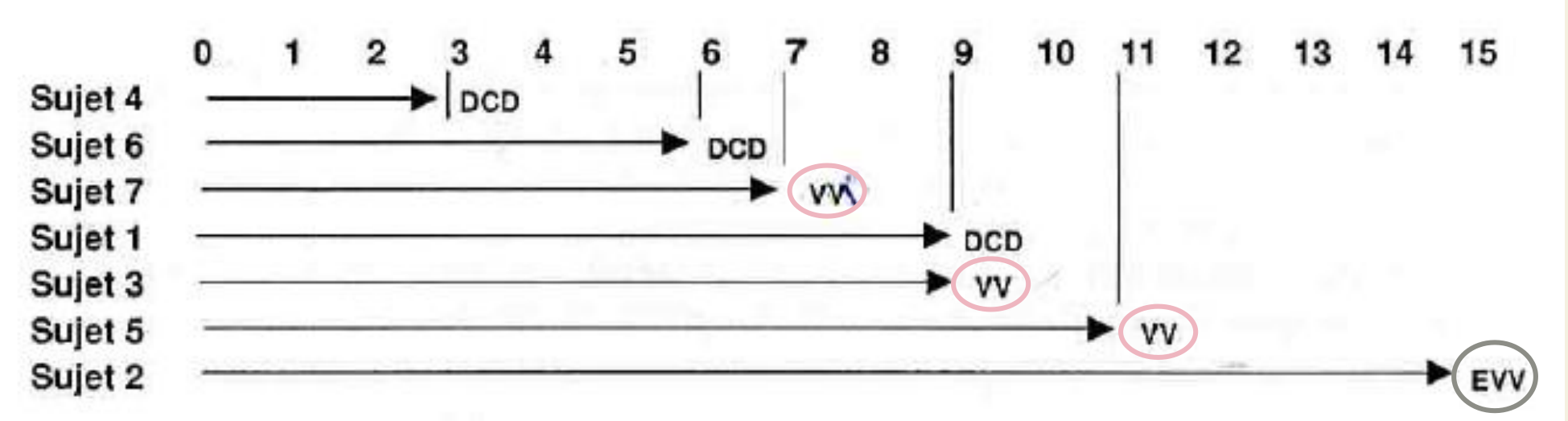
Récapitulatif des données

→ Date de Point (DP) = 01/06/1978

Données de base						
Sujet	Date Origine	Date des dernières nouvelles	Etat aux dernières nouvelles	Etat à la date de point	Temps de participation	Recul au 01/06/78
1	01/01/77	01/10/77	DCD	DCD	9	17
2	01/03/77	01/07/78	DCD	VV	15	15
3	01/05/77	01/02/78	VV	?	9	13
4	01/05/77	01/08/77	DCD	DCD	3	13
5	01/06/77	01/05/78	VV	?	11	12
6	01/07/77	01/01/78	DCD	DCD	6	11
7	01/08/77	01/03/78	VV	?	7	10

Estimation
Méthode de Kaplan-Meier

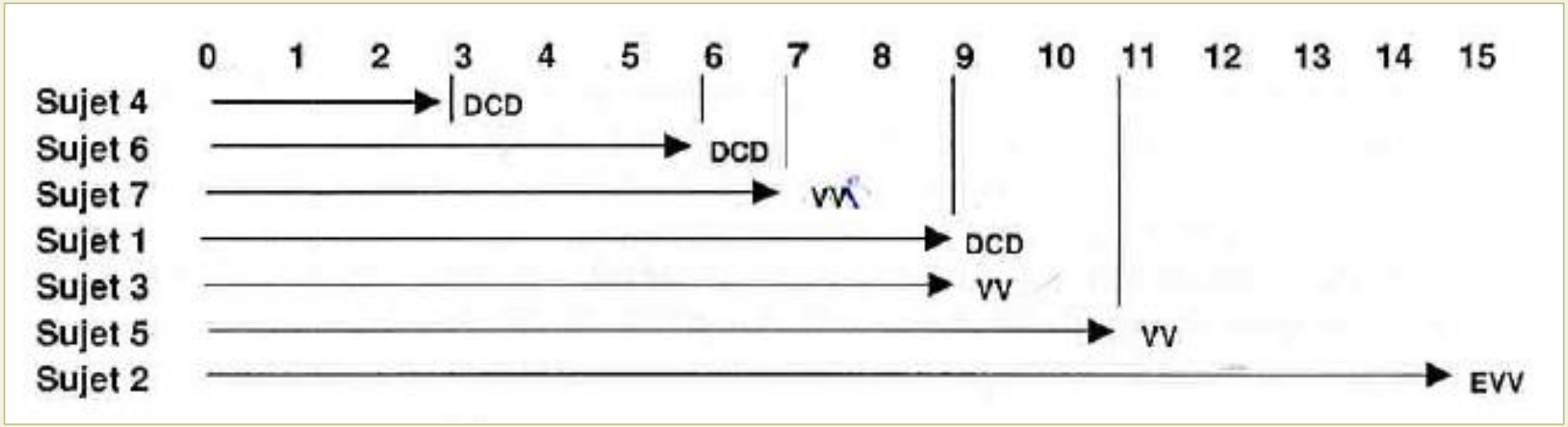
- *Classement des temps de participation par ordre croissant*



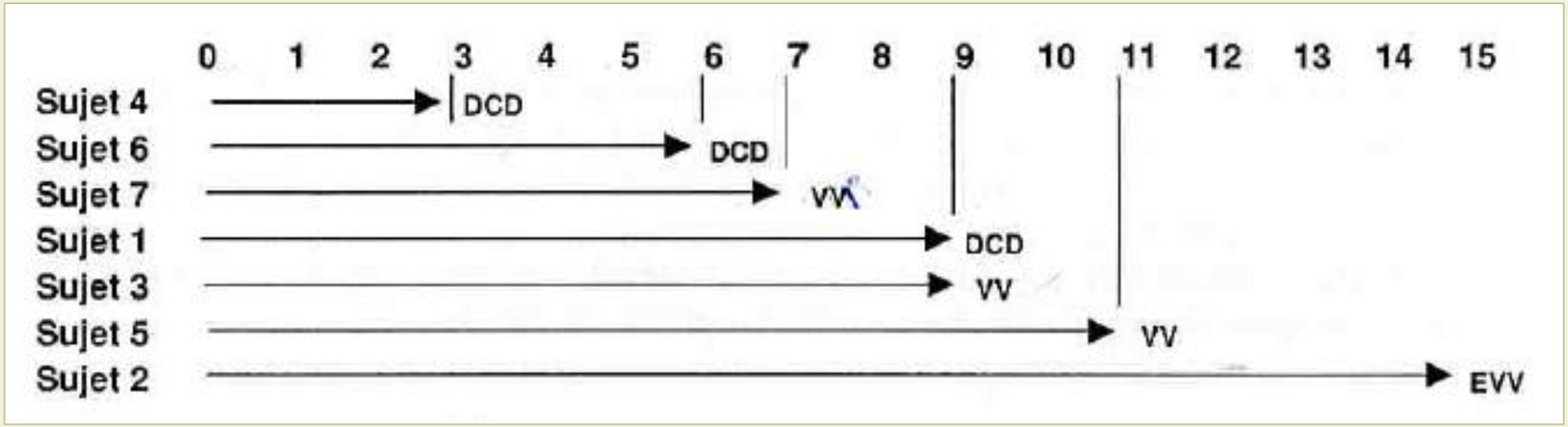
- Temps observés : 3, 6, 7*, 9, 9*, 11*, 15*
  - ⇨ *S(t) n'est estimée qu'aux temps avec événement : 3, 6, et 9*

t_i $(mois)$	n_i	m_i	c_i	q_i	S(t_i)
0	7	0	0	1	1

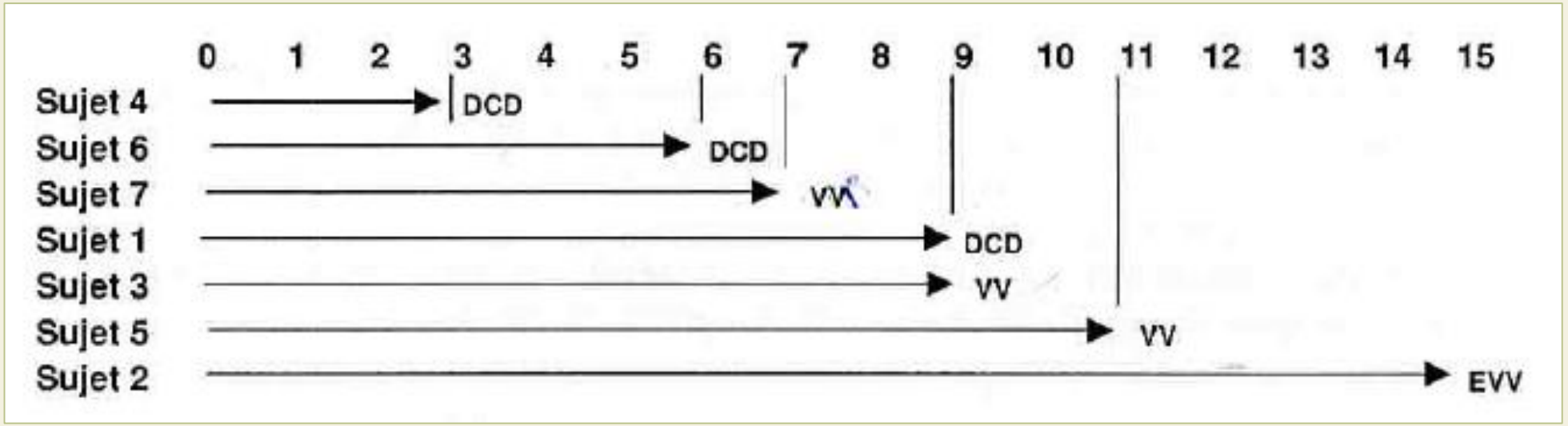
t_i : temps de participation du $i^{ème}$ événement (durée de survie)
n_i : nombre de personnes exposées au risque au temps t_i
m_i : nombre d'événements constatés au temps t_i
c_i : nombre de données censurées entre $[t_i ; t_{i+1}[$
q_i : probabilité de survie à chaque instant t_i
$S(t_i)$: probabilité cumulée de survie en t_i



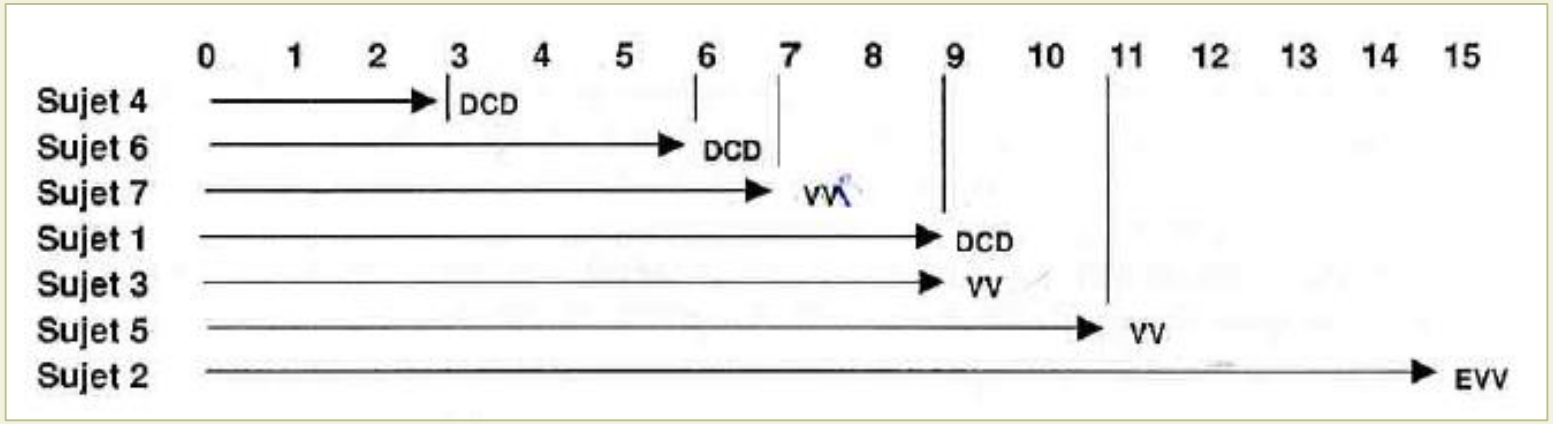
t_i $(mois)$	n_i	m_i	c_i	q_i	S(t_i)
0	7	0	0	1	1
3	$n_{t=3} = n_{t=0} - c_{[0;3[} = 7$	1	0	$n_{t=3} - m_{t=3} / n_{t=3} = {}^{6}/_{7}$	$q_{t=3} = 1 \times {}^{6}/_{7} = 0{,}857$



t_i $(mois)$	n_i	m_i	c_i	q_i	S(t_i)
0	7	0	0	1	1
3	$n_{t=3} = n_{t=0} - c_{[0;3[}$ $= 7$	1	0	$n_{t=3} - m_{t=3}/n_{t=3} = 6/7$	$q_{t=3} = 1 \times 6/7 = 0{,}857$
6	$n_{t=6} = n_{t=3} - m_{t=3} - c_{[3;6[}$ $= 7 - 1 - 0$ $= 6$	1	0	$n_{t=6} - m_{t=6}/n_{t=6} = 5/6$	$q_{t=3} \times q_{t=6} = 6/7 \times 5/6$ $= 5/7$ $= 0{,}714$

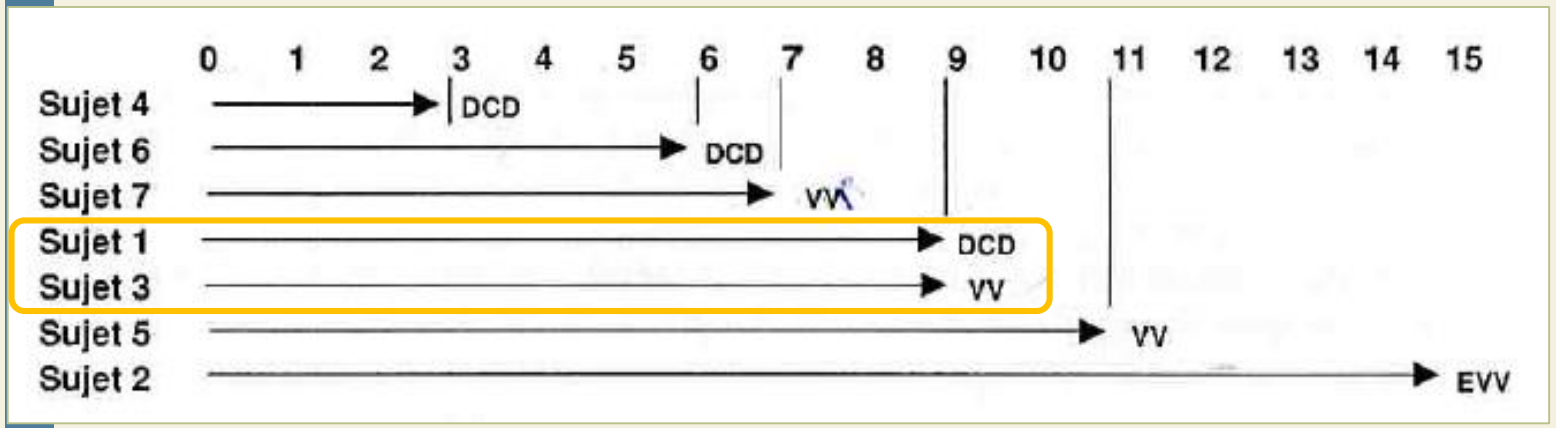


t_i $(mois)$	n_i	m_i	c_i	q_i	S(t_i)
0	7	0	0	1	1
3	$n_{t=3} = n_{t=0} - c_{[0;3[}$ $= 7$	1	0	$n_{t=3} - m_{t=3}/n_{t=3} = 6/7$	$q_{t=3}$
6	$n_{t=6} = n_{t=3} - m_{t=3} - c_{[3;6[}$ $= 7 - 1 - 0$ $= 6$	1	0	$n_{t=6} - m_{t=6}/n_{t=6} = 5/6$	$q_{t=3} \times q_{t=6} = 6/7 \times 5/6$ $= 5/7$ $= 0{,}857$
7	$n_{t=7} = n_{t=6} - m_{t=6} - c_{[6;7[}$ $= 6 - 1 - 0$ $= 5$	0	1	$n_{t=7} - m_{t=7}/n_{t=7} = 1$	



Calcul réalisé à titre d'illustration, mais non nécessaire : la fonction de survie n'est ré-estimée qu'aux temps avec événement

t_i $(mois)$	n_i	m_i	c_i	q_i	S(t_i)
0	7	0	0	1	1
3	$n_{t=3} = n_{t=0} - c_{[0;3[}$ $= 7$	1	0	$n_{t=3} - m_{t=3}/n_{t=3} = 6/7$	$q_{t=3}$
6	$n_{t=6} = n_{t=3} - m_{t=3} - c_{[3;6[}$ $= 7 - 1 - 0$ $= 6$	1	0	$n_{t=6} - m_{t=6}/n_{t=6} = 5/6$	$q_{t=3} \times q_{t=6} = 6/7 \times 5/6$ $= 5/7$ $= 0{,}857$
7	$n_{t=7} = n_{t=6} - m_{t=6} - c_{[6;7[}$ $= 6 - 1 - 0$ $= 5$	0	1	$n_{t=7} - m_{t=7}/n_{t=7} = 1$	



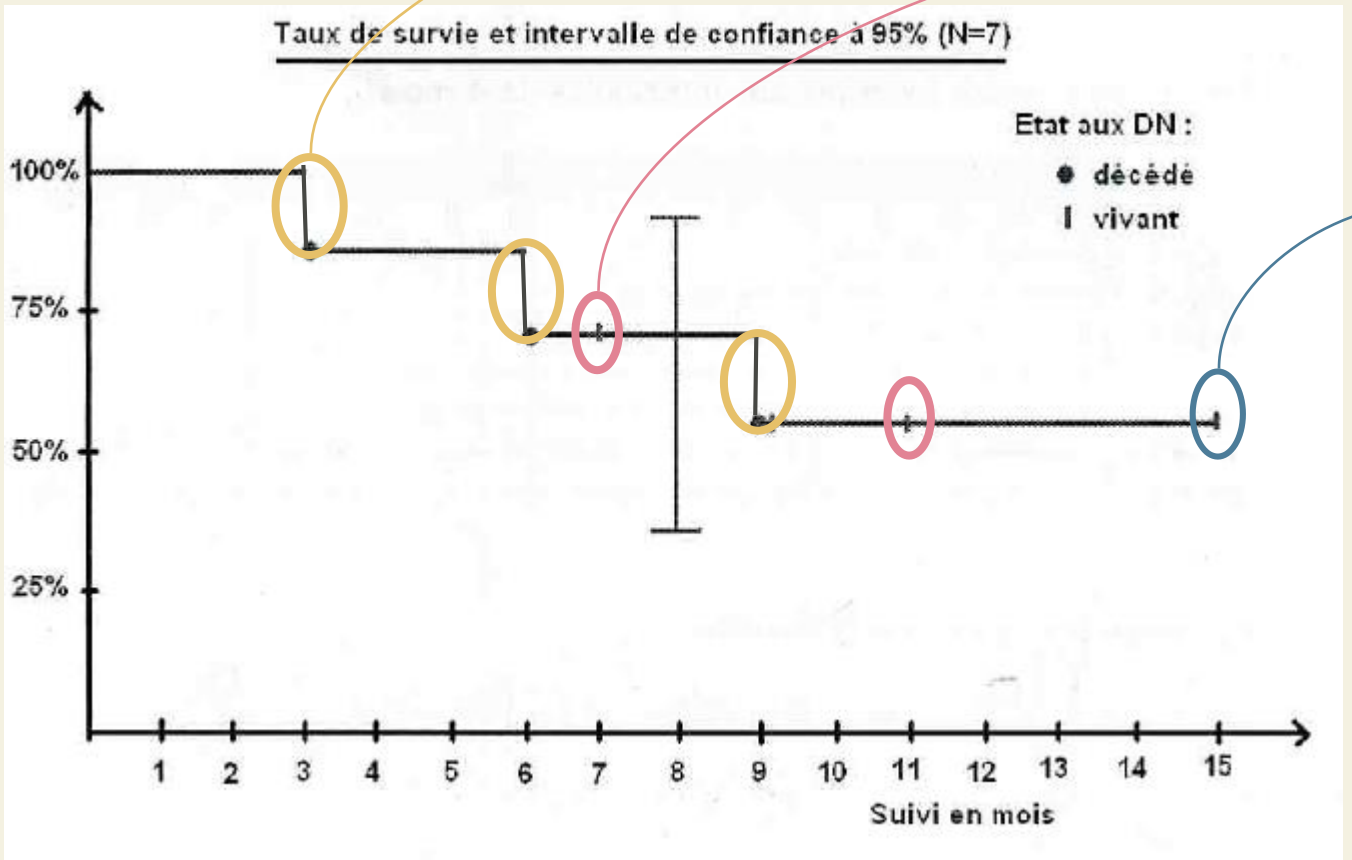
On considère que l'événement a eu lieu avant la censure
→ l'exclu-vivant est compté comme exposé

t_i $(mois)$	n_i	m_i	c_i	q_i	S(t_i)
0	7	0	0	1	1
3	$n_{t=3} = n_{t=0} - c_{[0;3[}$ $= 7$	1	0	$n_{t=3} - m_{t=3}/n_{t=3} = 6/7$	$q_{t=3}$
6	$n_{t=6} = n_{t=3} - m_{t=3} - c_{[3;6[}$ $= 7 - 1 - 0$ $= 6$	1	0	$n_{t=6} - m_{t=6}/n_{t=6} = 5/6$	$q_{t=3} \times q_{t=6} = 6/7 \times 5/6$ $= 5/7$ $= 0{,}857$
7	$n_{t=7} = n_{t=6} - m_{t=6} - c_{[6;7[}$ $= 6 - 1 - 0$ $= 5$	0	1	$n_{t=7} - m_{t=7}/n_{t=7} = 1$	
9	$n_{t=9} = n_{t=6} - m_{t=9} - c_{[6;9[}$ $= 6 - 1 - 1$ $= 4$	1	0	$n_{t=9} - m_{t=9}/n_{t=9} = 3/4$	$q_{t=7} \times q_{t=9} = 5/7 \times 3/4$ $= 15/28$ $= 0{,}536$

Courbe de Kaplan Meier

Evénement

Censure

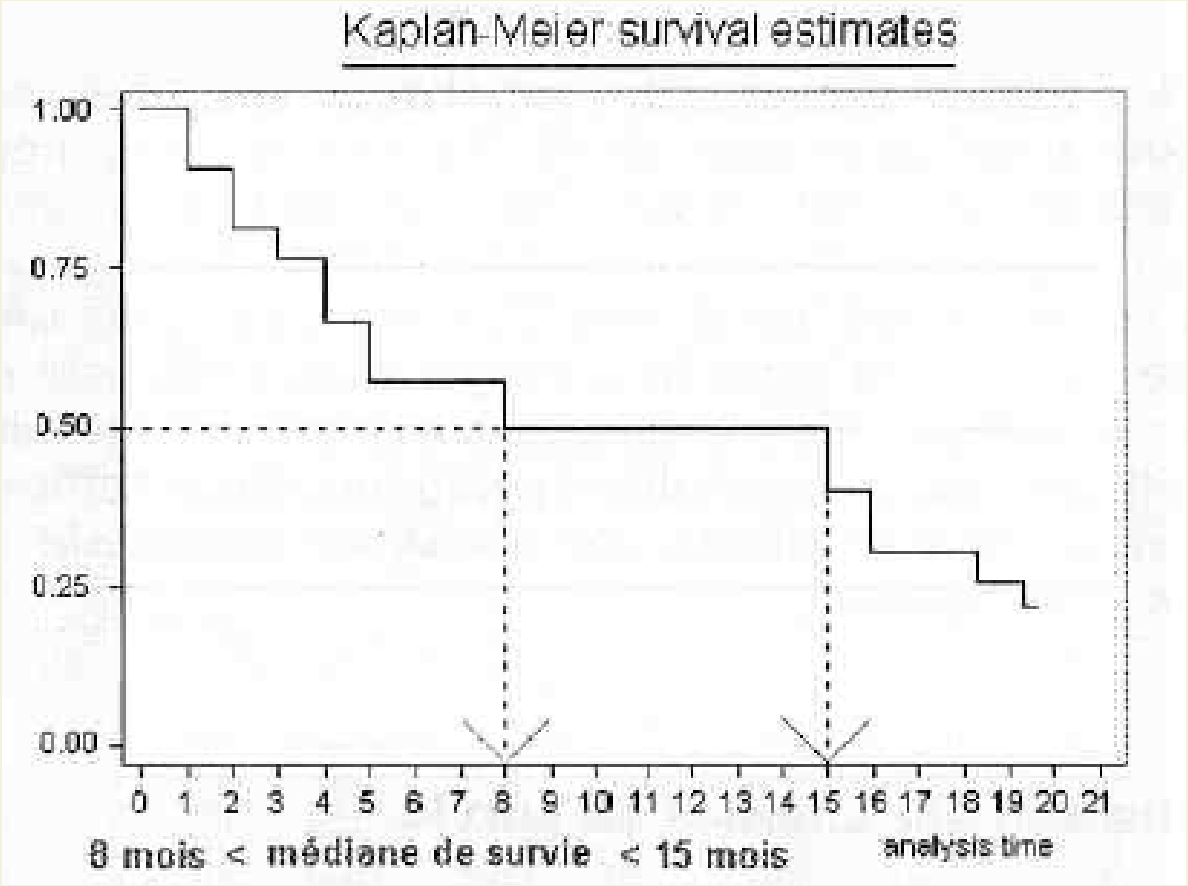


Courbe n'est pas définie au-delà de la dernière censure

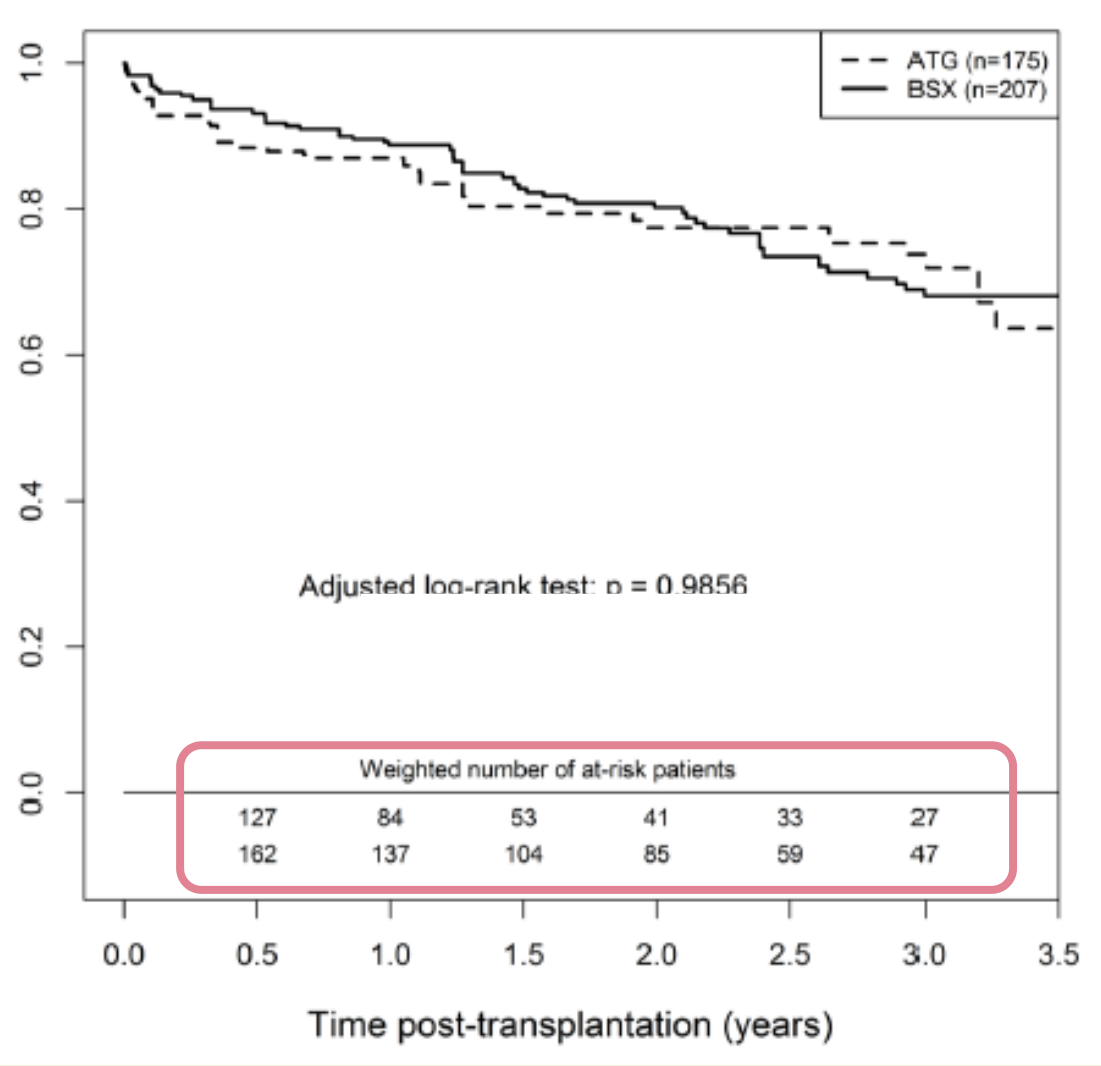
Ordonnée : taux de survie cumulatif

Abscisse : temps

Médiane de survie

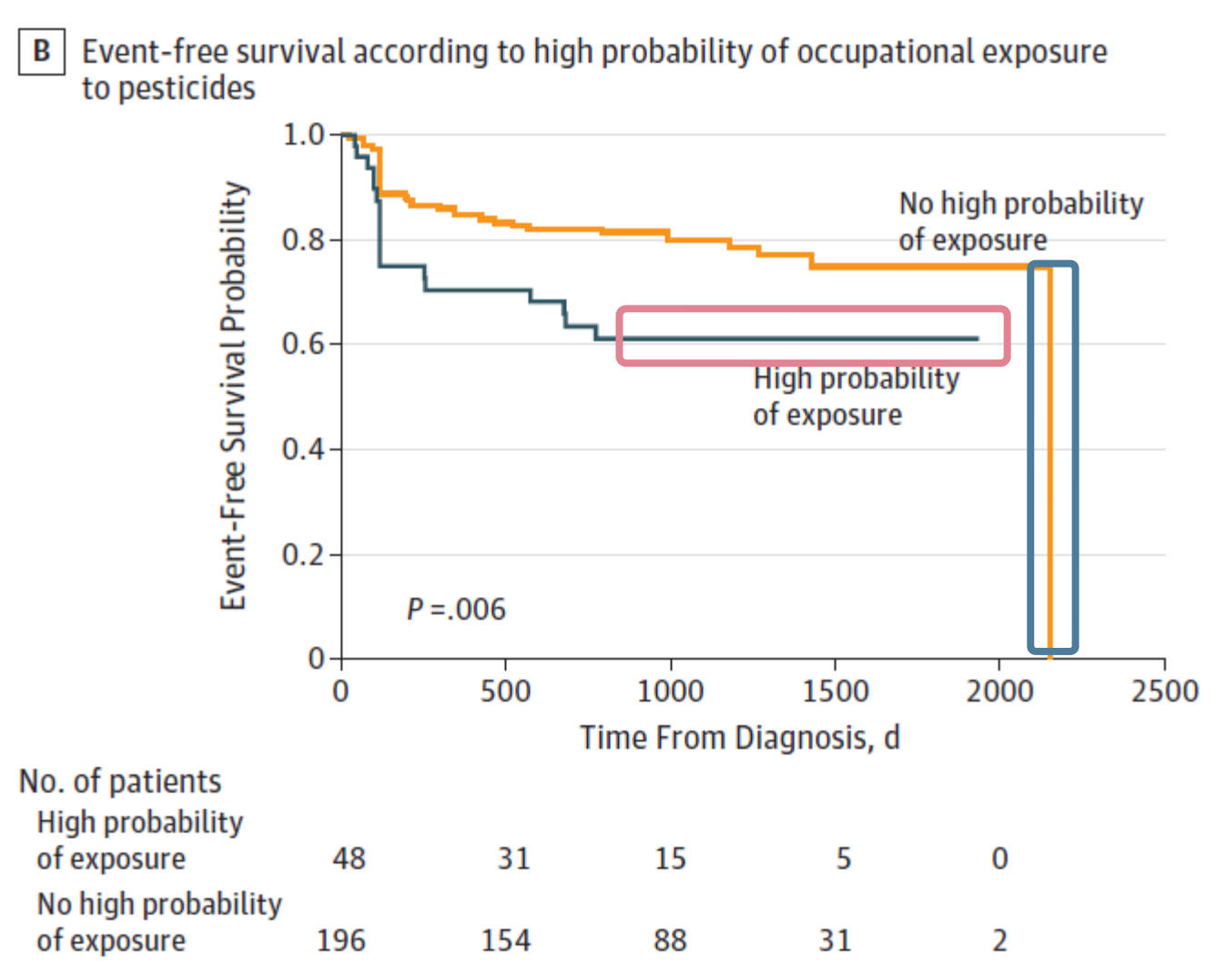


Nombre de sujets à risque

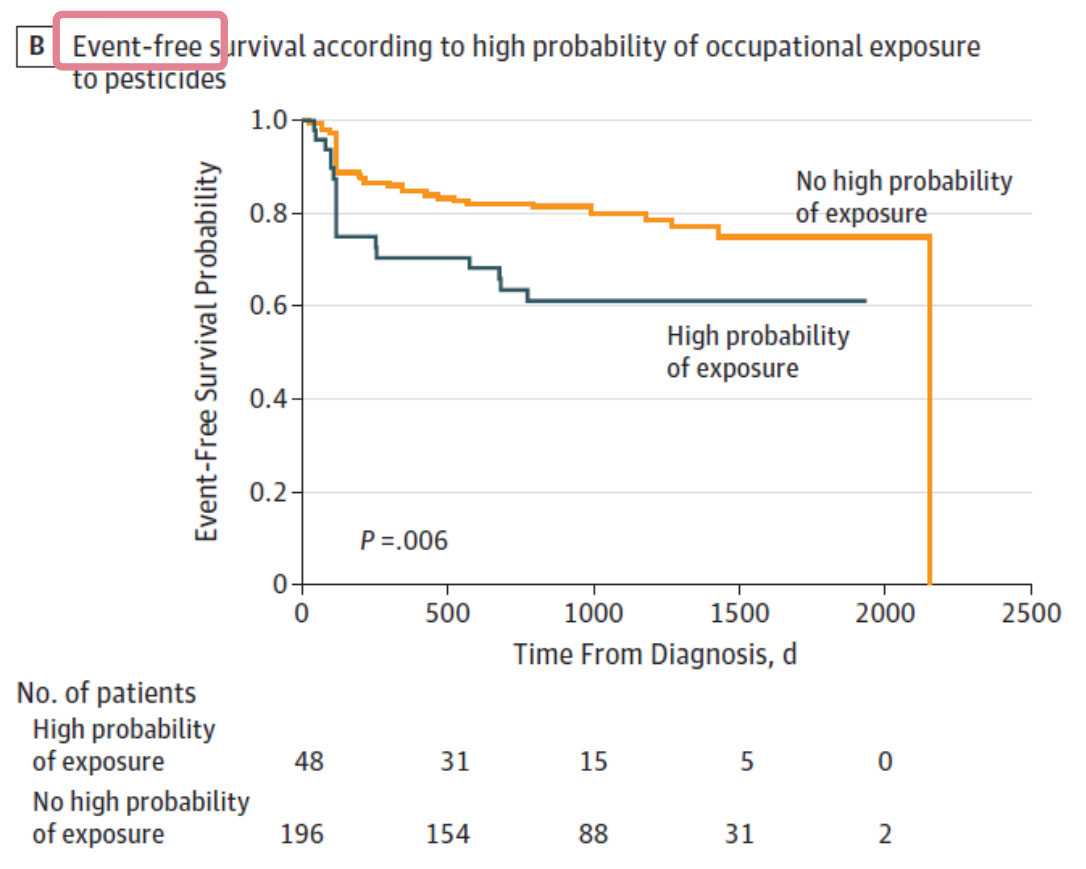


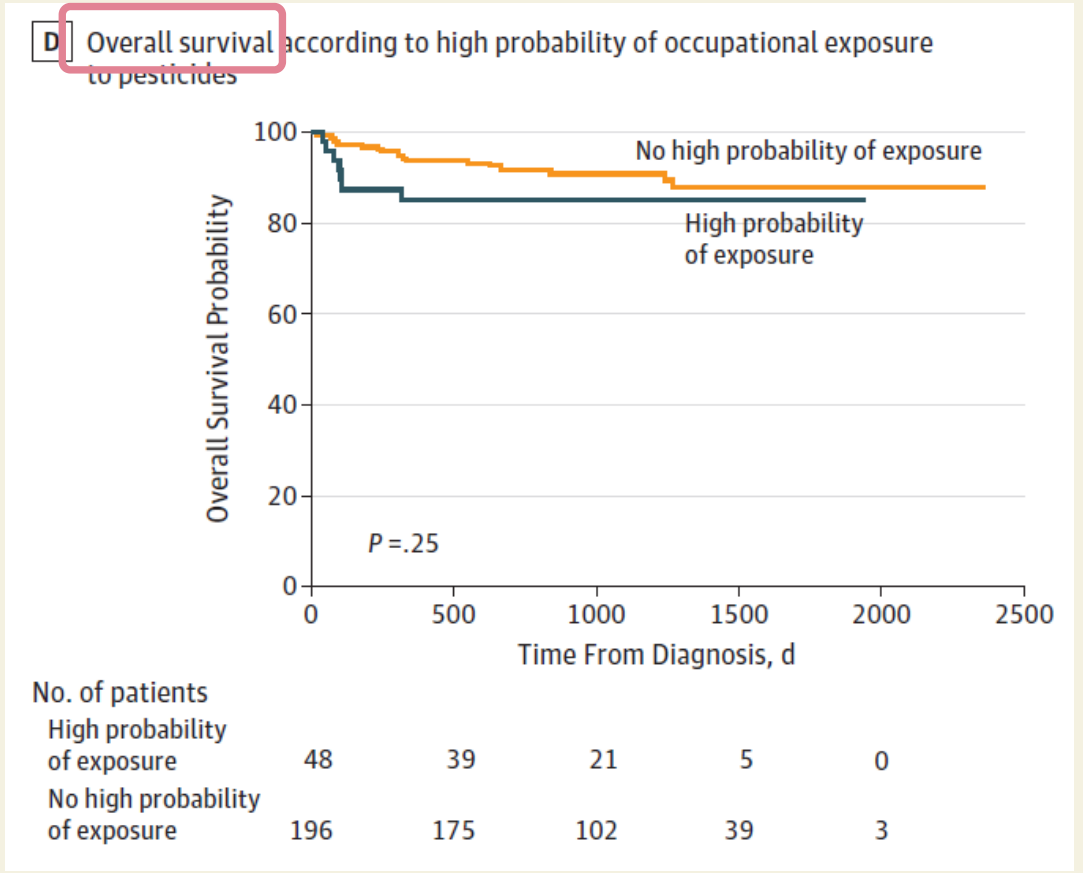
Nombre de sujets à risque au cours du temps pour chaque courbe de survie présentée

Interprétation



Interprétation





Traitement des censures

- Il faut vérifier que la censure est bien non informative
  → Mécanisme de censure ne doit pas être lié à la survenue de l'événement

    Exemple : les personnes sur le point de tomber malade ne se présentent pas au suivi

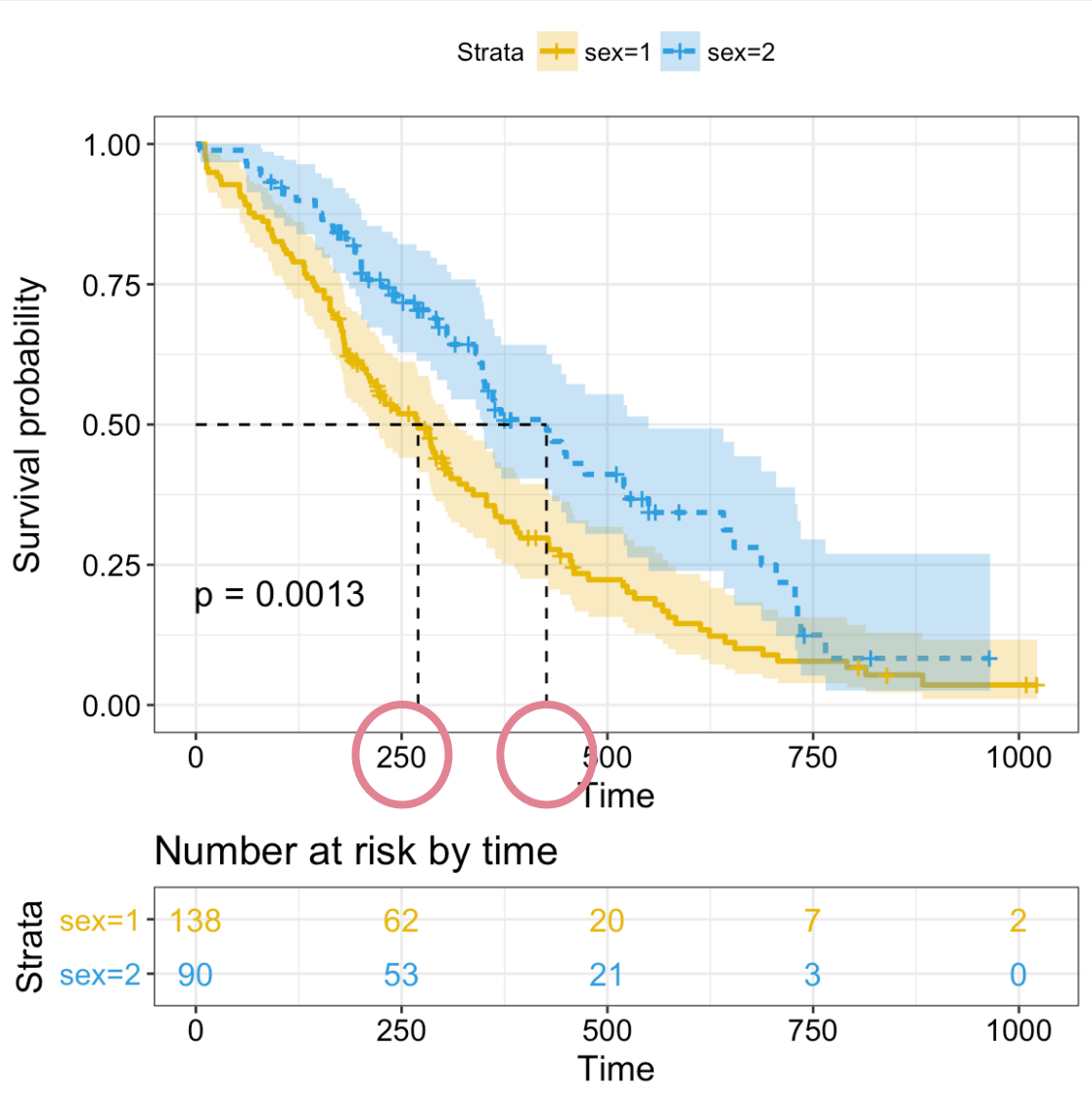
- Exclus-vivants : censure liée à la Date de Point (DP)
  →Donc forcément indépendante de la survenue de l'événement

- Perdus de vue : plusieurs aspects à vérifier !
  → Si peu nombreux (<10-15%) : assimilables aux Exclus-vivants
  → Si nombreux et comparables au reste de l'échantillon : assimilation aux Exclus-vivants
  →Si nombreux et non comparables au reste de l'échantillon : analyses de sensibilité

Exercice

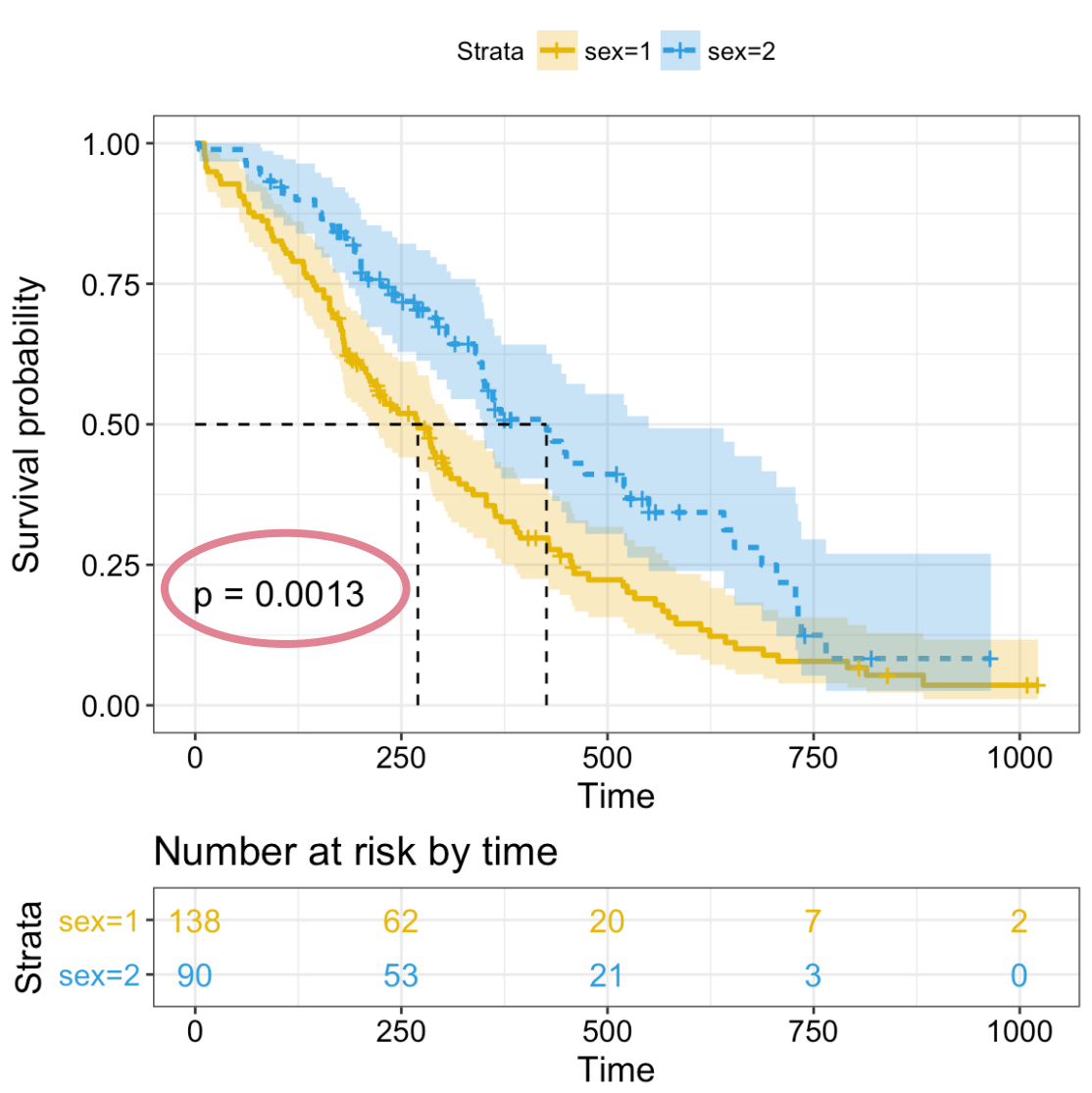
Survie de patients avec un cancer du poumon en fonction de leur sexe (1 = homme, 2 = femme)

Quelle est la médiane de survie pour chaque groupe ?



Source : Survival Analysis Basics, STHDA

Comparaison de courbes de survie



Test du logrank

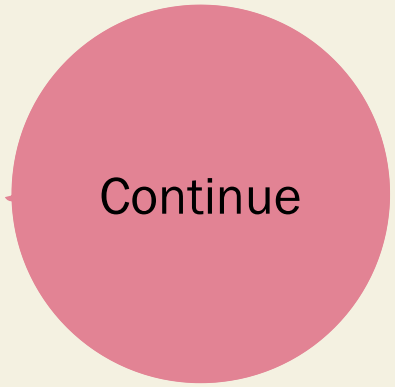
H_0 : les deux courbes de survie sont identiques (survie du groupe A = survie du groupe B)

H_1 : les deux courbes de survie sont différentes

Conditions d'application (hypothèses) du test du logrank

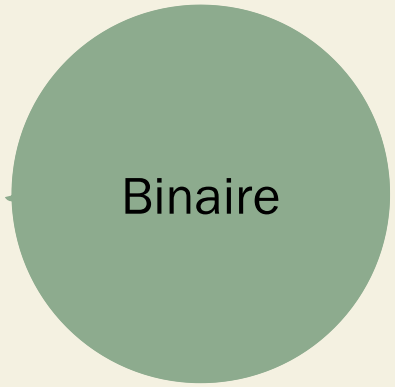
- Effectifs attendus > 5
- Censure non informative
- Risque dans un groupe est toujours supérieur à celui dans l'autre
- Les courbes ne se croisent pas

Ce que l'on va voir



Poids à la naissance

Régression linéaire



Survenue d'une maladie

Régression logistique

Délai jusqu'au décès

Kaplan Meier
Modèle de Cox

Modèle de Cox

Exprime le **risque instantané de survenue de l'événement** en fonction (d'un ou) de plusieurs facteurs explicatifs

→ **Hazard ratio** (HR) : risque relatif instantané d'événement

Univarié
(~test du logrank]

Multivarié

Exemple : Etude du risque de présenter un infarctus du myocarde au cours du suivi (exemple fictif)

Modèle de Cox

Variable	HR [IC 95%]	p-valeur
Sexe (F/H)	0,37 [0,25; 0,54]	0,004
Age (+10 ans)	1,22 [1,11; 1,35]	0,003
HTA : non	1	
HTA traitée	1,49 [1,22; 1,82]	0,004
HTA non traitée	2,23 [1,82; 2,72]	<0,001

Sexe : en moyenne, après ajustement sur les autres paramètres, le sexe féminin est associé à un risque significativement moins élevé de présenter un infarctus, avec un HR à 0,37 comparé au sexe masculin.

Modèle de Cox

Variable	HR [IC 95%]	p-valeur
Sexe (F/H)	0,37 [0,25; 0,54]	0,004
Age (+10 ans)	1,22 [1,11; 1,35]	0,003
HTA : non	1	
HTA traitée	1,49 [1,22; 1,82]	0,004
HTA non traitée	2,23 [1,82; 2,72]	<0,001

Age : en moyenne, après ajustement sur les autres paramètres, un âge plus avancé est associé à un risque significativement plus élevé de présenter un infarctus, avec un HR à 1,22 pour 10 ans.

Remarques

Hypothèse des risques proportionnels

→ Ratio constant au cours du temps

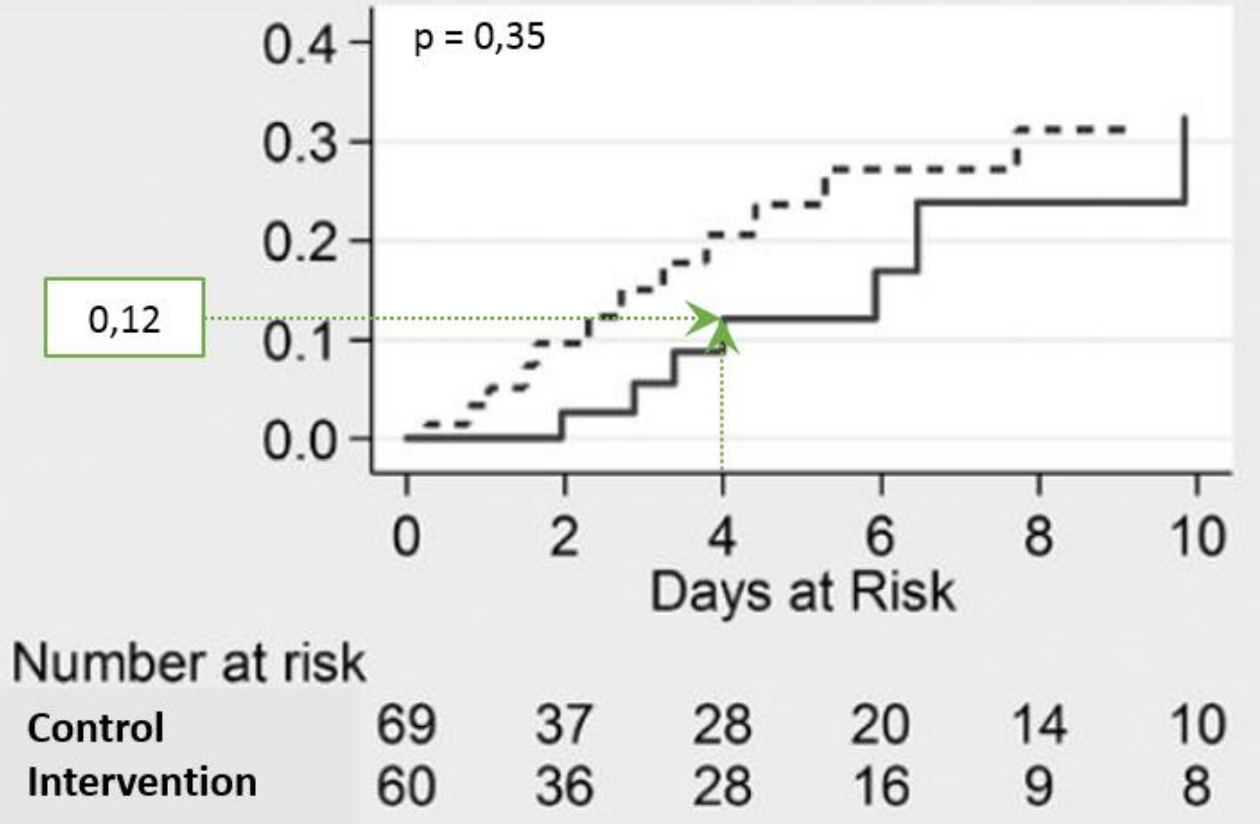
→ HR : rapport de risque instantané mais supposé constant dans le temps

Risque « au cours du suivi »

→ Si l'on compare le risque de décès dans deux groupes, la probabilité de décès sera égal à 1 si l'on attend suffisamment longtemps

Ne concerne pas que le décès

Remarques



Le « day at risk = 0 » correspond au jour de l'intervention chirurgicale. La courbe en trait plein représente le groupe avec douche préventive, la courbe en pointillés au groupe contrôle. L'événement recherché est la survenue d'infections nosocomiales.

- → Les analyses de survie ne concernent pas que les décès, tout événement pour lequel il est intéressant de considérer un délai d'apparition peut être étudié via ce type d'analyse
- → Les courbes de Kaplan-Meier peuvent être croissantes, dans ce cas, elles permettent alors d'estimer la probabilité de présenter l'événement considéré après un certain temps (ici la probabilité de présenter une infection nosocomiale dans les 4 jours après l'intervention dans le groupe avec douche préventive est de 12%).

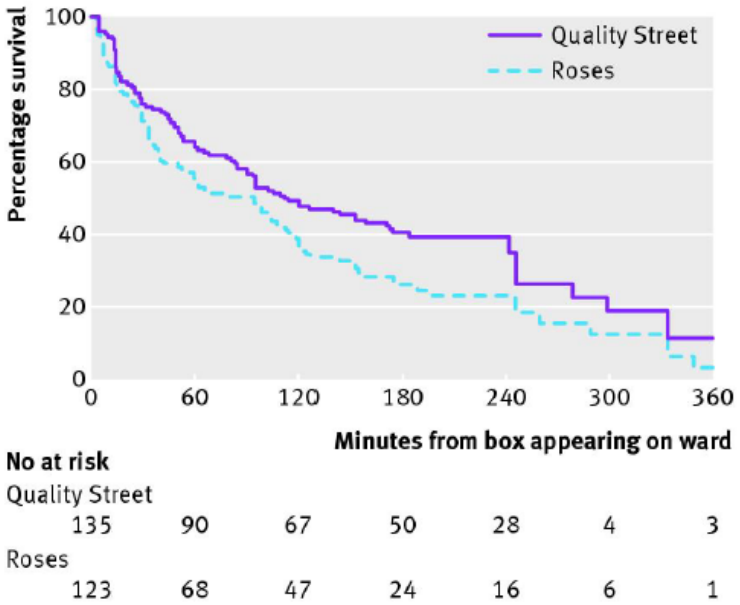
Exercice

Interpréter les résultats de cette étude : Gajendragadkar et al., The survival time of chocolates on hospital wards: covert observational study, *BMJ,* 2013

→ Numéro spécial de Noël du BMJ qui porte sur des études loufoques

→ Ici : comparaison de la durée de disparition des chocolats selon la marque, une fois les boîtes déposées dans les salles d'hôpital.

Résultat : préférence pour les Quality Streets par rapport aux Roses avec un HR de 0,70 statistiquement différent de 1.



Hazard ratio for survival of Roses v Quality Street 0.70, [95% CI : 0.53 - 0.93], p=0.014,

Fig 1 Kaplan-Meier survival curves for Quality Street and Roses chocolates across all wards