# Théorie

## Table of Contents



L'analyse des durées peut être vue comme l'étude d'une variable aléatoire $T$ qui décrit le temps d'attente jusqu'à l'occurence d'un évènement.

- La durée $T = 0$ est le début de l'exposition au risque (entrée dans le **Risk set**).
- $T$ est une mesure non négative de la durée.

## Le Risk Set

1. Il s'agit de la population "soumise" ou "exposée" au risque lorsque $T = t_i$.
2. Cette population varie dans le temps car:
   - Certaines personnes ont connu l'évènement, donc peuvent ne plus être soumises au risque (ex: décès si on analyse la mortalité).
   - Certaines personnes sortent de l'observation sans avoir (encore) observé l'évènement: décès si on analyse un autre type d'évènement, perdus de vue, fin de l'observation à une durée peu avancée dans un recueil rétrospectif.

**Exemples**:
* Les individus célibataires sont soumis au risque ……[remplir]

- Les individus mariés sont soumis au risque ……[remplir]
- Les individus au chômage sont soumis au risque ……..[remplir]
- Les individus qui travaillent sont soumis au risque …[remplir]

- Les individus vivants sont soumis au risque ……..[remplir]

## La Censure

Définition de la censure:
**Une observation est dite censurée lorsque la durée d'observation est inférieure à la durée d'exposition au risque**.

### Censure à droite

**Définition**
Certains individus n'auront pas (encore) connu l'évènement à la date de l'enquête après une certaine durée d'exposition. On a donc besoin d'un marqueur permettant de déterminer que les individus n'ont pas observé l'évènement sur la période d'étude.

**Pourquoi une information est-elle censurée (à droite)?**
- Fin de l'étude, date de l'enquête. - Perdu de vue, décès si autre évènement étudié.

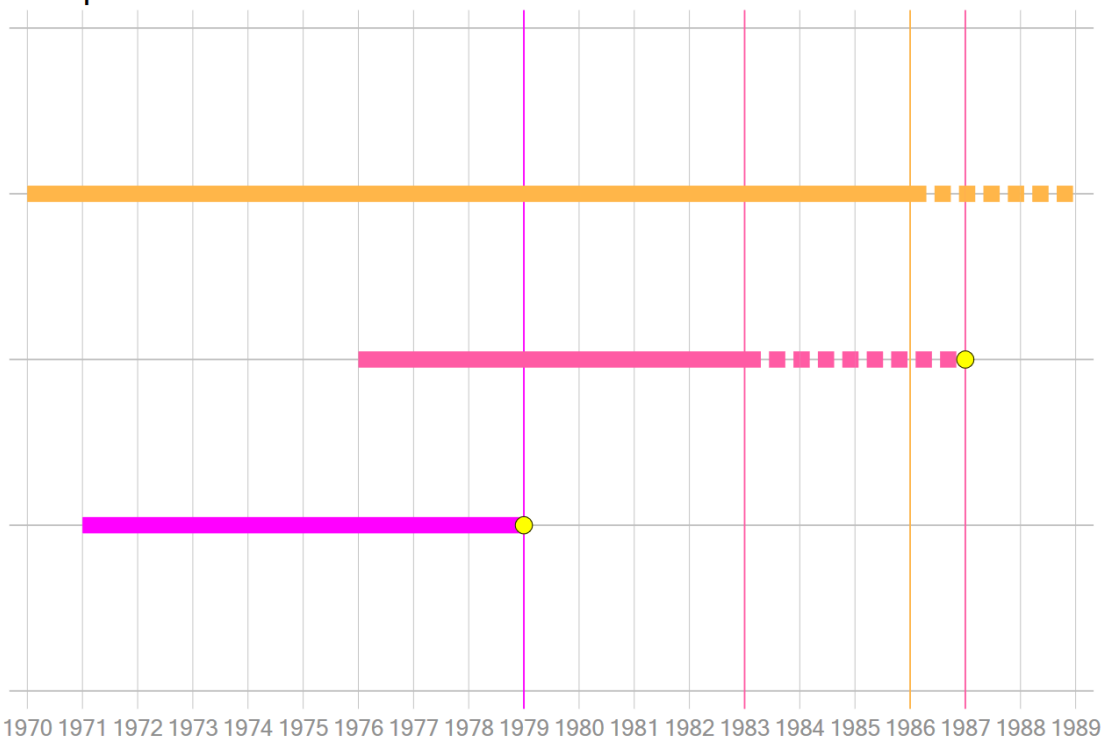
En pratique (important)

- **Ne pas exclure ces observations**, sinon on surestime la survenue de l'évènement.
- **Ne pas les considérer a-priori comme sorties sans connaître l'évènement**, elles peuvent connaître l'évènement après la date de l'enquête ou en étant perdues de vue. Sinon on sous estime la durée moyenne de survenue de l'évènement.

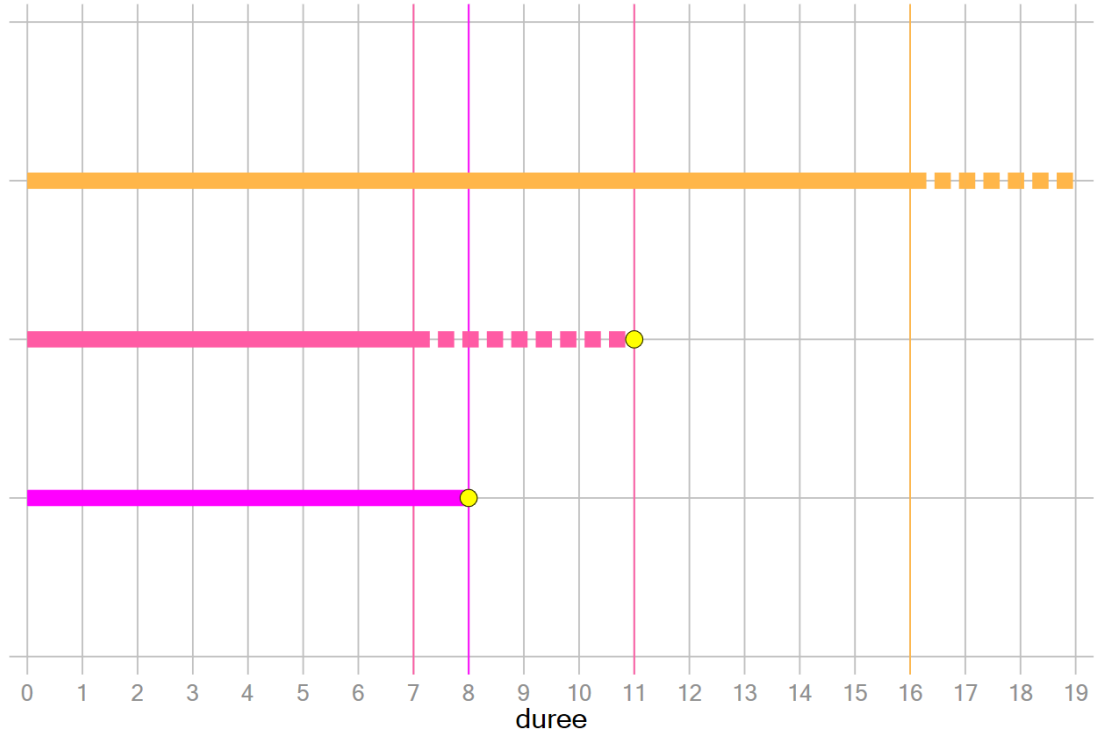
**Exemple**
On effectue une enquête auprès de femmes : On souhaite mesurer l'âge à la première naissance. Au moment de l'enquête, une femme est âgée de 29 ans et n'a pas (encore) d'enfant.
Cette information sera dite «censurée».
Elle est clairement encore soumise au risque après la date de l'enquête. Au niveau de l'analyse, elle sera soumise au risque à partir de ses premières règles jusqu'au moment de l'enquête.

**Hypothèse fondamentale**
Les observations censurées ont vis à vis du phénomène observé le même comportement que les observations non censurées. On dit que la **censure est non informative**. Elle ne dépend pas de l'évènement analysé.
*Exemple de censure informative*: analyse des décès avec un recueil prospectif. Si un individu est perdu de vue en raison d'une dégradation de son état de santé, l'indépendance entre la cause de la censure et le décès ne peut plus être assurée.

Temps calendaire




Durée



## Censure à gauche, troncature et censure par intervalle

**Censure à gauche**
L'évènement s'est produit avant le début période d'observation.
**Troncature à gauche (late-entry)**
Exemple: on analyse la survie d'une population. Seule la survie des individus vivants à l'inclusion peut-être analysée.
**Censure par intervalle** Un évènement peut avoir lieux entre 2 temps d'observations sans qu'on puisse les observer (ex: en criminologie récidive d'un delit entre deux arrestations).
Ces situations sont généralement plutôt bien contrôlées dans les recueils rétrospectifs. Elles sont assez courantes lorsque le recueil est de type prospectif.

# Les grandeurs

## Les grandeurs utilisées

La fonction de survie $S(t)$

La fonction de répartition $F(t)$

La fonction de densité $f(t)$

Le risque "instantané" $h(t)$

Le risque "instantané" cumulé $H(t)$

**Remarques**:

- Toutes ces grandeurs sont mathématiquement liées les unes par rapport aux autres. En connaître une permet d'obtenir les autres.

- Au niveau formel on se placera ici du point de vue où la durée mesurée est strictement continue. Cela se traduit, entre autre, par l'absence d'évènements dits "simultanés".

## La fonction de Survie $S(t)$

Dans ce type d'analyse, il est courant d'analyser la courbe de survie (ou de séjour).

**La fonction de survie donne la proportion de la population qui n'a pas encore connue l'évènement après une certaine durée** $t$. Elle y a "survécu".

Formellement, la fonction de survie est la probabilité de survivre au-delà de $t$, soit:

$$S(t) = P(T > t)$$

Propriétés: $S(0) = 1$ et $\lim_{t \to \infty} S(t) = 0$

## La fonction de répartition $F(t)$

C'est la probabilité de connaitre l'évènement jusqu'en $t$, soit:

$$F(t) = P(T \leq t)$$

$t \geq 0$ Soit $F(t) = 1 - S(t)$

Propriétés: $F(0) = 0$ et $\lim_{t \to \infty} F(t) = 1$

## La fonction de densité $f(t)$

- Pour une valeur de $t$ donnée, la fonction de densité de l'évènement donne la probabilité de connaitre l'évènement dans un petit intervalle de temps après $t$. Si $dt$ est proche de 0 alors cette probabilité tend également vers 0. On norme donc cette probabilité par $dt$.
- En temps continu, la fonction de densité est donnée par la dérivée de la fonction de répartition: $f(t) = F'(t) = -S'(t)$. Formellement la fonction de densité $f(t)$ s'écrit:

$$f(t) = \lim_{dt=0} \frac{P(t \leq T < t + dt)}{dt}$$

## Le risque instantané $h(t)$

Concept fondamental de l'analyse des durées:

$$h(t) = \lim_{dt=0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}$$

- $P(t \leq T < t + dt | T \geq t)$ donne la probabilité de survenue de l'évènement sur l'intervalle $[t, t + dt[$ *conditionnellement à la survie au temps* $t$.

- La quantité obtenue donne un nombre moyen d'évènements que connaîtrait un individu durant une unité de temps choisie.

On peut écrire également: $h(t) = \frac{f(t)}{S(t)} = \frac{F\prime(t)}{S(t)} = -\frac{S\prime(t)}{S(t)}$

## Le risque cumulé $H(t)$

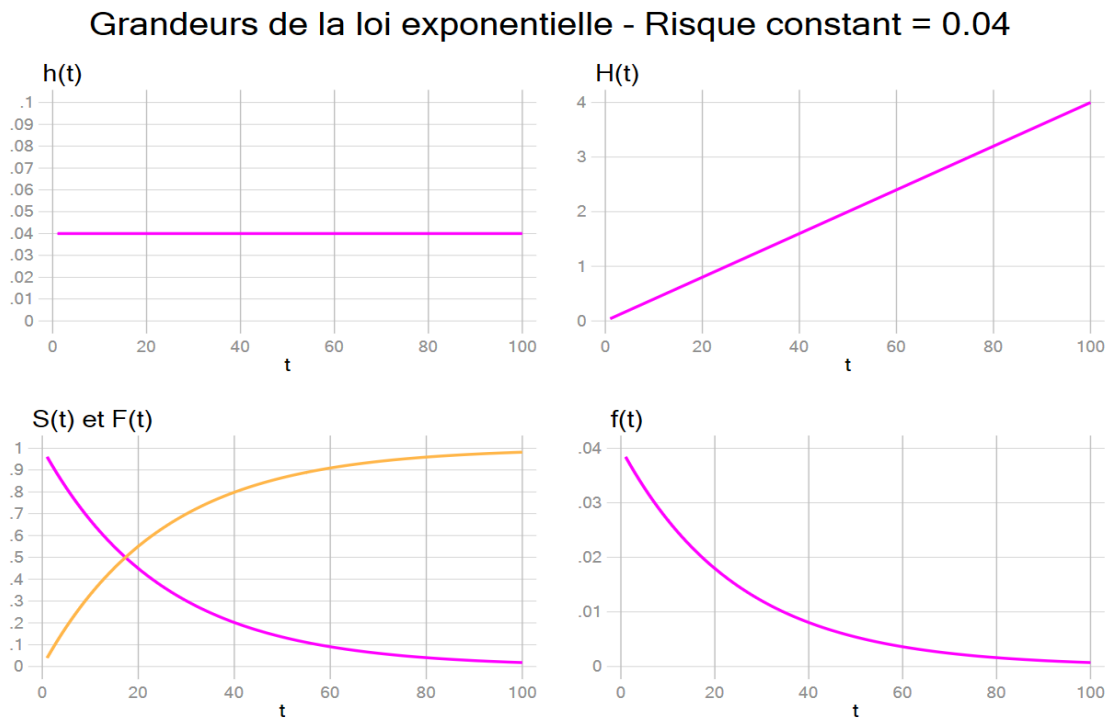
Le risque cumulé est égal à : $H(t) = \int_0^t h\,(u)du = -log(S(t))$

On peut alors le réécrire toutes les autres quantités:

- $S(t) = e^{-H(t)}$
- $F(t) = 1 - e^{-H(t)}$
- $f(t) = h(t) \times e^{-H(t)}$

Si on pose que le risque est strictement constant au cours du temps: $h(t) = a$ (on parle de **loi exponentielle** - cf partie sur les modèles AFT):

- $h(t) = a$
- $H(t) = a \times t$
- $S(t) = e^{-a \times t}$
- $F(t) = 1 - e^{-a \times t}$
- $f(t) = a \times e^{-a \times t}$




**Application: risque et échelles temporelles**:
Fortement inspiré de l'*excellent* cours de **Gilbert Colletaz**: https://www.univ-orleans.fr/deg/masters/ESA/GC/sources/Econometrie%20des%20Donnees%20de%20Survie.pdf

1. Durant les mois d'hiver, entre le 1er janvier et le 1er avril (3 mois), la probabilité d'attraper un rhume chaque mois est de 48% (il s'agit bien d'un risque). Quelle est le risque d'attraper le rhume durant la saison froide?
   $h(t) = \frac{0.48}{1/3} = 1.44$. On peut donc s'attendre a attraper 1.44 rhume durant la période d'hiver.

2. On passe une année en vacances dans une région où la probabilité de décéder chaque mois est évaluée à 33%. Quelle est le risque de décéder pendant cette année sabbatique? $h(t) = \frac{0.33}{1/12} = 3.96$

*Remarque*: le risque peut donc être supérieur à 1. En soit cela ne pose pas de problème comme il s'agit d'un nombre moyen d'évènements espérés (exemple: "taux" de fécondité), mais pour des évènements qui ne peuvent pas se répéter, évènements dits "absorbants", l'interprétation n'est pas très intuitive.
On peut donc prendre l'inverse du risque qui mesure la durée moyenne (espérée) jusqu'à l'occurence de l'évènement.
On retrouve donc un concept classique en analyse démographique comme l'espérance de vie (survie): la question n'est pas de savoir si "on" va mourir ou non, le risque

indépendement du temps étant par définition égal à 1, mais jusqu'à quand on peut espérer survivre.
- Pour le rhume, la durée moyenne est de $1/1.44 = 0.69$ du trimestre hivernal, soit approximativement le début du mois de mars. - Pour l'année sabbatique, la durée moyenne de survie (l'espérance de vie) est de $1/3.96 = 0.25$ d'une année soit 3 mois après l'arrivée dans la région.

**Exercice**

- On a une population de 100 cochons d'Inde.
- On analyse leur mortalité (naturelle).

- Ici l'analyse est en temps discret.
- La durée représente le nombre d'année de vie.
- Il n'y a pas de censure à droite.

| Durée | Nombre de décès |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 3 |
| 4 | 9 |
| 5 | 30 |
| 6 | 40 |
| 7 | 10 |
| 8 | 3 |
| 9 | 2 |
| 10 | 1 |
| N=100 | |

A quel âge le risque de mourir des cochons d'Inde est-il le plus élevé? Quelle est la valeur de ce risque?