Analyse de Survie

Professeur Abdellatif El Afia

Rappel

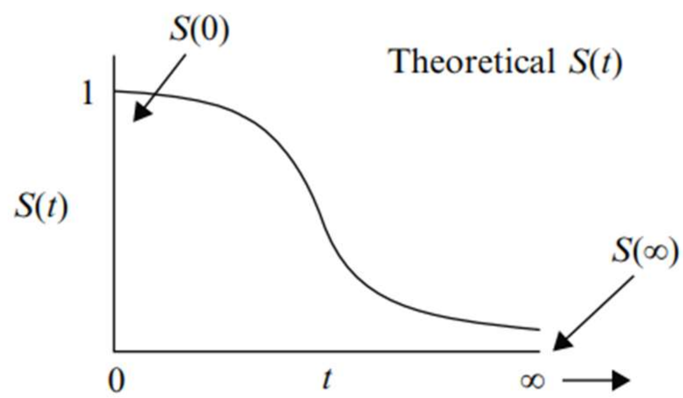
- L'analyse de survie est un ensemble de procédures statistiques pour l'analyse de données dans lesquelles la variable de résultat d'intérêt est le **temps** jusqu'à ce qu'un **événement** se produise.
- Modèle : Exposition, Evénement, Temps, Variables X.
- La censure : on ne connait pas le temps exactement.
- Censure à droite $\Rightarrow$ temps de survie observé < temps réel de survie.
- Fonctions de survie :
  - $\boldsymbol{S(t)} = P(T \geq t) = 1 - F(t),\ t > 0$
  - $\boldsymbol{F(t)} = 1 - S(t)$
  - $\boldsymbol{f(t)} = -\frac{d}{dt}S(t)$
  - $\boldsymbol{\lambda\ (t)} = h(x) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}$
  - $\boldsymbol{\Lambda(t) = H(t)} = \int_0^t h(x)dx\ ,\ t \geq 0$

Travail 1 :

- Trouver des exemples d’analyses de survie.
- Modéliser ces exemples (Exposition, Evénement, Temps, Variables X.).
- Trouver une data de survie

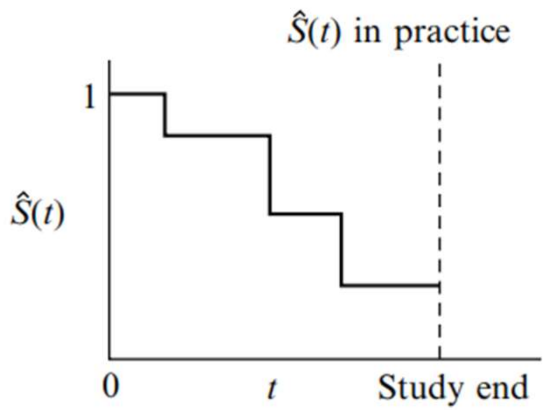
Fonction de survie

Théoriquement, comme t varie de 0 à l'infini, la fonction de survie ***S(t)*** est représentée graphiquement sous la forme d'une courbe lisse décroissante, de la valeur ***S(t)=1*** pour ***t=0,*** et tend vers 0 quand t tend vers l'infini.

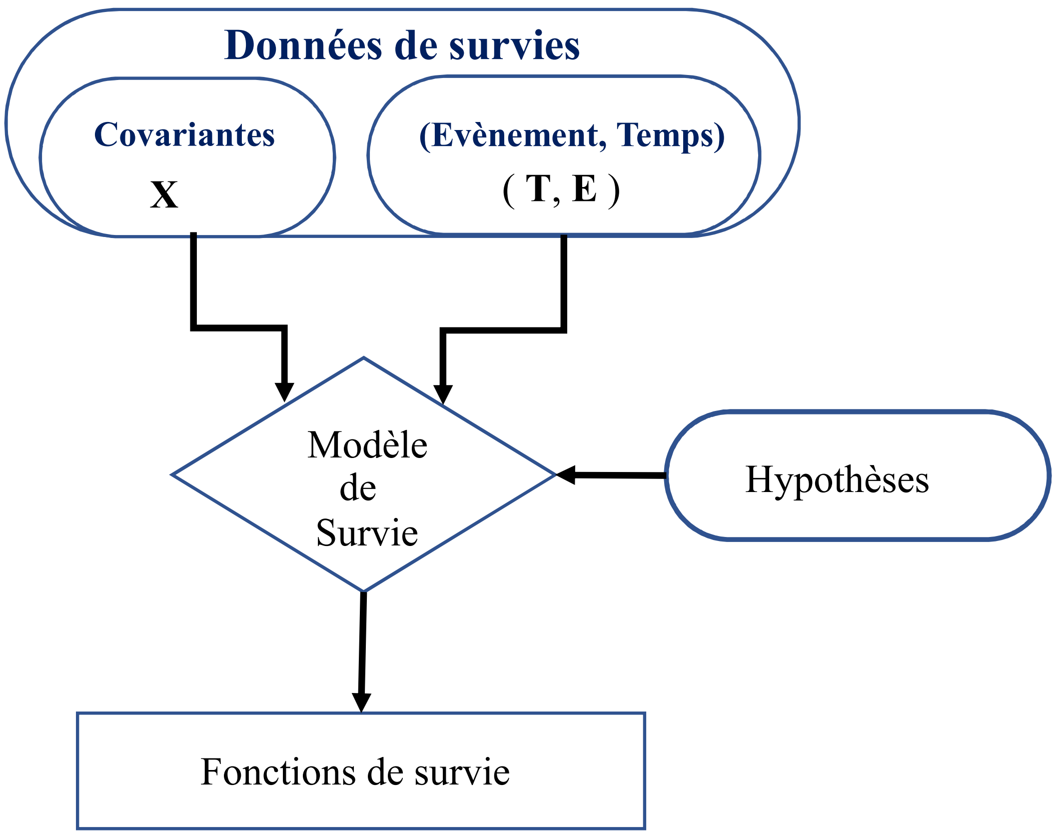


Fonction de survie

En pratique, en utilisant des données, les courbes de survie estimées sont généralement des fonctions en escalier.



Modèle de survie



Fonction de survie :

$$S(t) = P(T \geq t) = 1 - F(t),\ t > 0$$

Fonction de densité :

$$f(t) = -\frac{d}{dt}S(t)$$

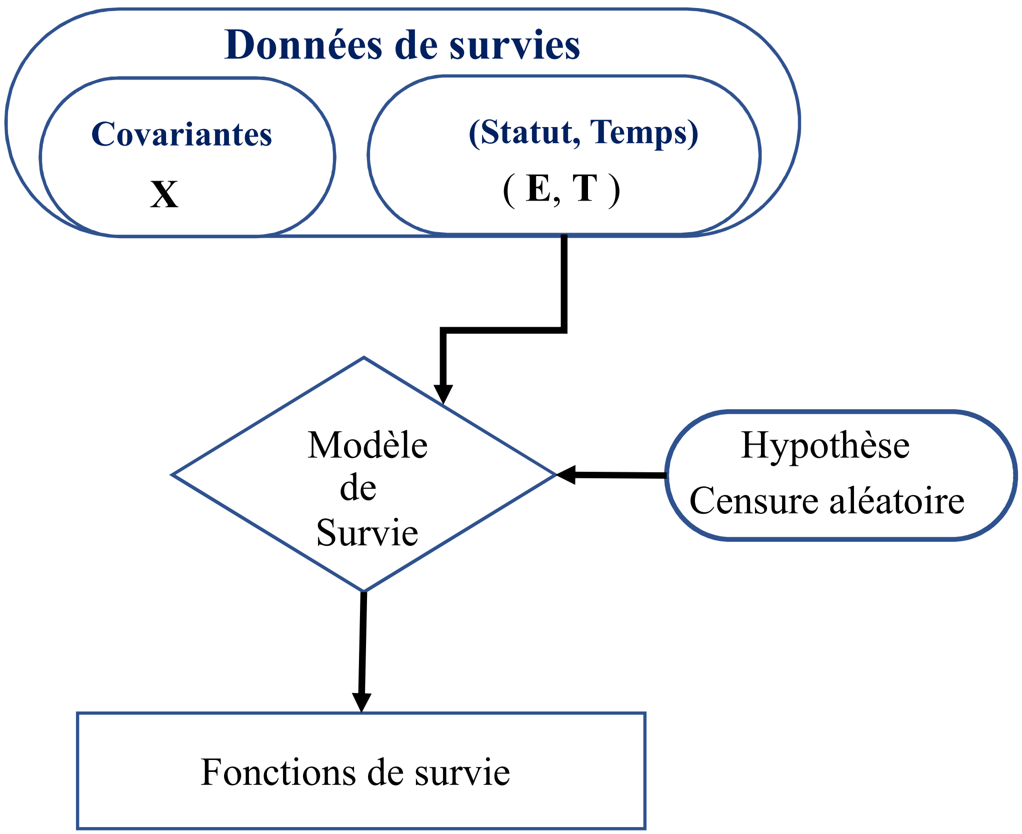
Fonction de risque :

$$\lambda(t)\Delta t \approx P(t \leq T < t + \Delta t | T \geq t)$$

Fonction de risque Cumulative :

$$\Lambda(t) = \int_0^t \lambda(x)dx\ ,\ t \geq 0$$

Modèles Non Paramétriques



- **Estimateur de Kaplan Meier (KM) :**

$$\hat{S}(t_k) = \prod_{t_k<t} S(t_{k-1})\left(1 - \frac{d_k}{n_k}\right) \qquad 1 < k < j$$

- **Estimateur de Nelson–Aalen :**

$$\widetilde{H}(t) = \sum_{t_i<t} \frac{d_i}{n_i}$$

- **Tables de survie.**

Estimateur de Kaplan Meier (KM)

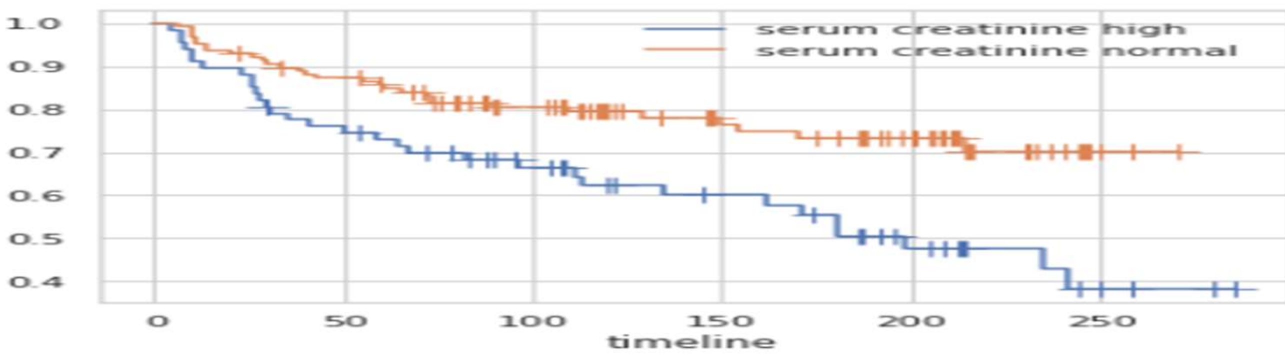
Hypothèses :

- Censure non-informative.
- Temps de survie indépendants.

▪l'hypothèse nulle: H_0:"$Il\ n'\ y\ a\ pas\ de\ difference\ statistique\ entre\ les\ deux\ groupes$"

Estime : la fonction de survie ***S(t)*** en fonction en éscaliers**.**

Utilisé souvent pour mesurer la fraction d'individus en vie pour une certaine durée, et comparer la survie de deux ou plusieurs groupes.



Kaplan-Meier étapes

- Classer les temps de survie dans un ordre croissant commençant par la valeur 0.

- La deuxième colonne indique la fréquence des événements à chaque instant d'événement distinct.
- La troisième colonne donne la fréquence des personnes censurées, noté qf, dans l'intervalle de temps commençant par l'instant d'événement t(f) jusqu'à l'instant d'événement suivant mais non compris, désigné par t(f+1).
- La dernière colonne donne les individus au risque, qui ont survécu au moins jusqu'au temps t(f).
- On calcule la probabilité de survie ***S(t)*** de chaque temps de survie prenant en compte la censure si elle existe.

Exemple pratique

The data: remission times (weeks) for two groups of leukemia patients

Group 1 ($n = 21$) treatment	Group 2 ($n = 21$) placebo
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+,	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Note: + denotes censored

	# failed	# censored	Total
Group 1	9	12	21
Group 2	21	0	21

Descriptive statistics:

$$\bar{T}_1 \text{(ignoring + 's)} = 17.1, \quad \bar{T}_2 = 8.6$$

$$\bar{h}_1 = .025, \quad \bar{h}_2 = .115, \quad \frac{\bar{h}_2}{\bar{h}_1} = 4.6$$

Exemple pratique

Group 2 ($n = 21$) placebo
1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Group 2 (placebo)

$t_{(f)}$	n_f	m_f	q_f
0	21	0	0
1	21	2	0
2	19	2	0
3	17	1	0
4	16	2	0
5	14	2	0
8	12	4	0
11	8	2	0
12	6	2	0
15	4	1	0
17	3	1	0
22	2	1	0
23	1	1	0

Exemple pratique

Group 1 ($n = 21$) treatment
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+,

Group 1 (treatment)

$t_{(f)}$	n_f	m_f	q_f
0	21	0	0
6	21	3	1
7	17	1	1
10	15	1	2
13	12	1	0
16	11	1	3
22	7	1	0
23	6	1	5
>23	—	—	—

KM formule = Produit limite

$$\hat{S}(t_k) = \hat{S}(t_{k-1}) \times \hat{P}(T > t_k \mid T \geq t_k)$$

$$\hat{S}(t_k) = \prod_{i=1}^{k} \hat{P}\,(T > t_i \mid T \geq t_i)$$

$$P(A \cap B) = P(A) \times P(B \mid A)$$

$$\hat{S}(t_k) = \prod_{t_k < t} S(t_{k-1}) \left(1 - \frac{m_k}{n_k}\right) \qquad 1 < k < j$$

Log-Rank test

- Un test chi-deux χ^2 utilisant un grand échantillon pour fournir une comparaison globale de KM courbes.

- Utilisé pour tester l'hypothèse nulle:

$$H_0 : "Il\ n'y\ a\ pas\ de\ difference\ statistique\ entre\ les\ deux\ groupe\text{«}$$

- Utilise la technique d'observés Vs attendues sur les catégories, ces dernières étant définies par chacun des évènements ordonnés pour l'ensemble des données analysées (les deux groupes).

Log-Rank test

- Attendues (expected) :
  - $e_i^1 = ecpected_i^{group\ 1} = \frac{At\ risk_i^{group1}}{At\ risk_i^{group1} + At\ risk_i^{group2}} \times \left(Failure_i^{group1} + Failure_i^{group2}\right)$

- Observées (observed) :
  - $o_i^1 = observed_i^{group1} = Failure_i^{group1}$
- Obseerved – expected :
  - $O^g - E^g = \sum_i (o_i^g - e_i^g)$

Log-Rank test

$$\textit{Log-rank statistic} = \frac{(O^g - E^g)^2}{Var(O^g - E^g)}$$

- Log-rank statistic est approximativement du khi-deux χ^2 avec un degré de liberté.
- Approximation : $\chi^2 = \sum_g \frac{(O^g - E^g)^2}{E^g}$
- La p-value de Log-rank détermine si H_0 est rejetée ou non.

Log-Rank test

EXAMPLE

Remission data: $n = 42$

	# failures		# in risk set	
$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
4	0	2	21	16
5	0	2	21	14
6	3	0	21	12
7	1	0	17	12
8	0	4	16	12
10	1	0	15	8
11	0	2	13	8
12	0	2	12	6
13	1	0	12	4
15	0	1	11	4
16	1	0	11	3
17	0	1	10	3
22	1	1	7	2
23	1	1	6	1

Expected cell counts:

$$e_{1f} = \underbrace{\left(\frac{n_{1f}}{n_{1f} + n_{2f}}\right)}_{\text{Proportion in risk set}} \times \underbrace{\left(m_{1f} + m_{2f}\right)}_{\text{\# of failures over both groups}}$$

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f} + n_{2f}}\right) \times \left(m_{1f} + m_{2f}\right)$$

Log-Rank test

EXAMPLE

Expanded Table (Remission Data)

		# failures		# in risk set		# expected		Observed-expected	
f	$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}	e_{1f}	e_{2f}	$m_{1f}-e_{1f}$	$m_{2f}-e_{2f}$
1	1	0	2	21	21	$(21/42) \times 2$	$(21/42) \times 2$	−1.00	1.00
2	2	0	2	21	19	$(21/40) \times 2$	$(19/40) \times 2$	−1.05	1.05
3	3	0	1	21	17	$(21/38) \times 1$	$(17/38) \times 1$	−0.55	0.55
4	4	0	2	21	16	$(21/37) \times 2$	$(16/37) \times 2$	−1.14	1.14
5	5	0	2	21	14	$(21/35) \times 2$	$(14/35) \times 2$	−1.20	1.20
6	6	3	0	21	12	$(21/33) \times 3$	$(12/33) \times 3$	1.09	−1.09
7	7	1	0	17	12	$(17/29) \times 1$	$(12/29) \times 1$	0.41	−0.41
8	8	0	4	16	12	$(16/28) \times 4$	$(12/28) \times 4$	−2.29	2.29
9	10	1	0	15	8	$(15/23) \times 1$	$(8/23) \times 1$	0.35	−0.35
10	11	0	2	13	8	$(13/21) \times 2$	$(8/21) \times 2$	−1.24	1.24
11	12	0	2	12	6	$(12/18) \times 2$	$(6/18) \times 2$	−1.33	1.33
12	13	1	0	12	4	$(12/16) \times 1$	$(4/16) \times 1$	0.25	−0.25
13	15	0	1	11	4	$(11/15) \times 1$	$(4/15) \times 1$	−0.73	0.73
14	16	1	0	11	3	$(11/14) \times 1$	$(3/14) \times 1$	0.21	−0.21
15	17	0	1	10	3	$(10/13) \times 1$	$(3/13) \times 1$	−0.77	0.77
16	22	1	1	7	2	$(7/9) \times 2$	$(2/9) \times 2$	−0.56	0.56
17	23	1	1	6	1	$(6/7) \times 2$	$(1/7) \times 2$	−0.71	0.71
Totals		9	21			19.26	10.74	−10.26	−10.26

Log-Rank test

of failure times

$$O_i - E_i = \sum_{f=1}^{\overset{\text{\# of failure times}}{\downarrow}\atop 17} \left(m_{if} - e_{if} \right),$$

$$i = 1, 2$$

$$\mathrm{Var}(O_i - E_i) = \sum_j \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}$$

$$i = 1,\ 2$$

EXAMPLE

$O_1 - E_1 = -10.26$
$O_2 - E_2 = 10.26$

EXAMPLE

Using Stata: Remission Data

Group	Events observed	Events expected
1	9	19.25
2	21	10.75
Total	30	30.00

Log rank = chi2(2) = 16.79
P-Value = Pr > chi2 = 0.000

Two groups:

$O_2 - E_2$ = summed observed minus expected score for group 2

$$\text{Log-rank statistic} = \frac{(O_2 - E_2)^2}{\mathrm{Var}(O_2 - E_2)}$$

Pysurvival python

```
[ ] import io
    col_names=['age','anaemia','creatinine','diabetes','ejection_fraction','high_blood_pressure','platelets','serum_creatinine','serum_sodium','sex','smoking','time','DE
    data= pd.read_csv(io.BytesIO(uploaded['heart_failure_clinical_records_dataset.csv']))
    data.head(5)
```

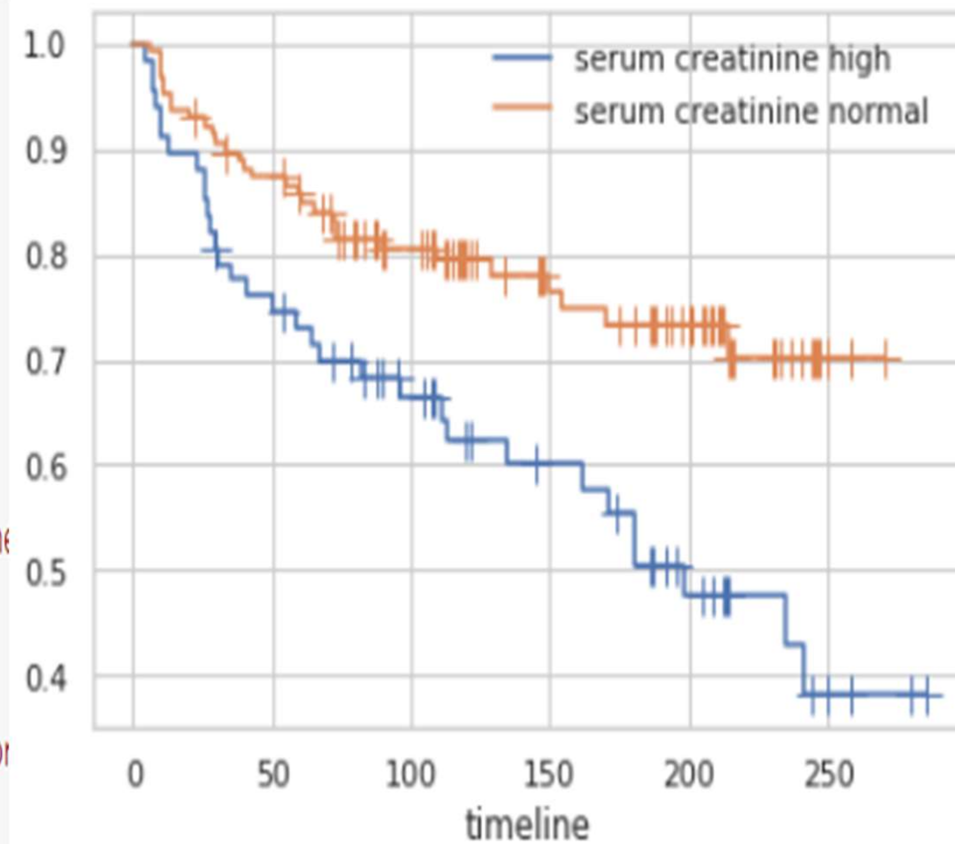
	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4	1
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7	1
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7	1
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8	1

Pysurvival KM

```
T = male['time']
E = male['DEATH_EVENT']
group = male['serum_creatinine']
eleveted_creatinine = (group > 1.2  )
normal_creatinine = (group <= 1.2)


kmf.fit(T[eleveted_creatinine], E[eleveted_creatinine], label='serum creatinine
ax = kmf.plot(ci_show=False, show_censors=True)


kmf.fit(T[normal_creatinine], E[normal_creatinine], label='serum creatinine nor
ax = kmf.plot(ci_show=False, show_censors=True)
```



Lifelines Log-Rank test

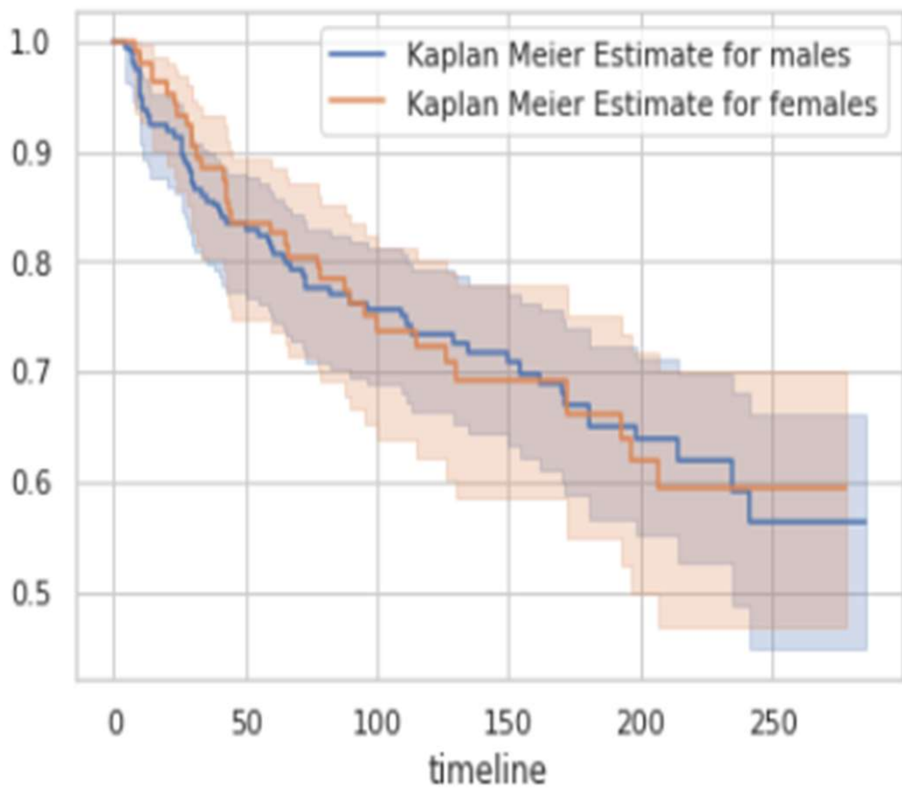
```
from lifelines import KaplanMeierFitter
from lifelines.statistics import logrank_test
```

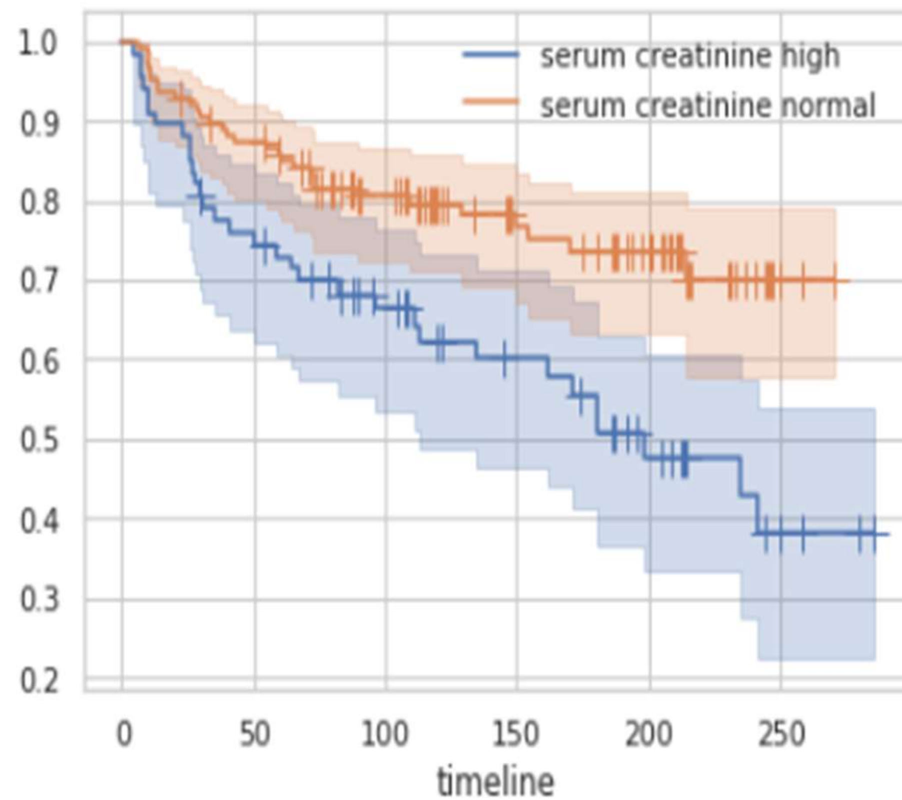
```
results=logrank_test(T[normal_creatinine],
                     T[eleveted_creatinine],
                     event_observed_A=E[normal_creatinine]
                     event_observed_B=E[eleveted_creatinin
print(results)
```

```
<lifelines.StatisticalResult: logrank_test>
               t_0 = -1
 null_distribution = chi squared
degrees_of_freedom = 1
         test_name = logrank_test

---
 test_statistic      p  -log2(p)
           9.79 <0.005      9.16
```

KM avec Intervalle de confiance





Log-Rank test

- pour genre

```
<lifelines.StatisticalResult: logrank_test>
               t_0 = -1
 null_distribution = chi squared
degrees_of_freedom = 1
         test_name = logrank_test

---
 test_statistic    p  -log2(p)
           0.00 0.95      0.07
```

- Pour serum creatinine

```
<lifelines.StatisticalResult: logrank_test>
               t_0 = -1
 null_distribution = chi squared
degrees_of_freedom = 1
         test_name = logrank_test

---
 test_statistic      p  -log2(p)
           9.79 <0.005      9.16
```

Plus de tests

- Wilcoxen
- Tarone-Ware
- Peto
- Flemington-Harrington

Kaplan-Meier

• Avantages

- Simple à interpréter.
- Permet d'estimer $S(t)$.

• Inconvénients

- Pas de formule-Fonction.
- Ne permet pas d'estimer le « Hazard ratio »
- Que quelques catégoriques X

Autres Modèles Non-paramètriques

- Estimateur de **Breslow** du risque cumulé.
- Estimateur de **Nelson-Aalen** du risque cumulé.
- Estimateur de **Harrington-Fleming** du risque cumulé.
- Estimation des variances.
- Life tables.

TP 2 :

- Subdiviser la data en deux groupes (ou plus) selon une variable.
- Implémenter KM pour les deux groupes et tracer KM courbe.
- Tester l'hypothèse nulle par le Log-Rank test
- Comparer vos résultats avec ceux trouver par les bibliothèques existantes.