

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Les méthodes d'analyse de données de survie

- Les méthodes de comparaison de courbes de Survie
- Les modèles non-paramétriques
  - Le Log-Rank Test
  - Exemple : taux de gestation – Vêlage 2 ans
  - Log-Rank test : la statistique de test
  - Log-Rank test : Conditions d'utilisation
  - Autres développements autour du test du Log-Rank
- Les modèles semi-paramétriques
  - Les modèles à risques proportionnels
  - Le modèle de Cox
  - Résultats : conduite Vêlage 2 ans
  - Adéquation du modèle de Cox

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Les méthodes d'analyse de données de survie

- Les modèles paramétriques
  - Les modèles à temps accélérés
  - Le modèle de Weibull
  - Différentes lois de distributions de fonctions de Risques
  - Une particularité du modèle de Weibull
  - Exemple des pneumonies de veaux
  - Adéquation du modèle de Weibull

### Les méthodes de comparaison de courbes de Survie

Comparaisons globales des survies dans des populations (facteurs de risque)

Cadre : recherche de facteurs de risque de survenue d'un événement, la « mort »

Trois catégories de méthodes :

#### 1. Les méthodes non-paramétriques

on compare les populations sur les occurrences de mort.

On ne pose aucun modèle de loi sur la répartition des occurrences en fonction du temps. On se contente d'utiliser l'ordre d'apparition des événements.

#### 2. Les méthodes semi-paramétriques

on compare des risques d'incidences selon les populations .

On pose un modèle multiplicatif sur l'évolution des risques en fonction de l'exposition à des facteurs de risques. On ne modélise pas l'évolution des risques en fonction du temps (modèles à « risques proportionnels »). On se contente d'utiliser l'ordre d'apparition des événements (modèle de Cox, 1972).

### Les méthodes de comparaison de courbes de Survie

#### 3. Les méthodes paramétriques

On compare les temps de survie selon les populations.

On pose des modèles de distributions des temps de survie conditionnelles aux facteurs de risques définissant les populations.

« modèles à temps accélérés » (***Accelerated Failure Time Models***)

Avantage : ces modèles utilisent la durée des intervalles de temps entre deux évènements contrairement aux méthodes non-paramétriques et semi-paramétriques qui n'utilisent que les rang des évènements.

Inconvénient : les résultats sont dépendants des lois posés sur les temps...

*Peu utilisé en Epidémiologie (à ma connaissance...)*

*Très large utilisation du modèle de Cox (trop?)*

# LES MÉTHODES NON PARAMETRIQUES

## Le Log-Rank Test

### Les méthodes non-paramétriques

Plusieurs tests sont proposés :

Test de Wilcoxon, test de Tarone-Ware, Test Prentice-Peto-Peto, test de Harrington-Flemmings et test du Log-Rank.

Ils supposent tous que les temps de survenus des morts sont fixés (raisonnement conditionnels aux évènements, « *sachant que l'on a observé ces évènements dans cet ordre ...* »)

Ils se différencient par l'importance qu'ils accordent à différentes parties des courbes de Survie.

Les tests de Wilcoxon et Tarone-Ware pondèrent les statistiques calculées à chaque temps où survient au moins un événement par le nombre d'individus à risque avant l'événement. Les tests de Prentice-Peto-Peto et de Harrington-Flemmings pondèrent respectivement par la Survie, une fonction de la survie  $S(t)$ .

Le test du log-Rank accorde la même importance à tous les temps (poids=1).

---

### Les méthodes non-paramétriques

#### Le test du Log-Rank

L'idée (*cas de la comparaison de deux courbes de Survie*) :

1. On dresse une table de contingence à chaque temps  $t$  où se produit au moins un événement ('mort').

On calcule le nombre de mort(s) attendu(s) à  $t$  dans chacun des deux groupes sous l'hypothèse ( $H_0$ ) que le risque de mort est le même dans les deux groupes.

2. On calcule, pour chaque groupe, l'écart  $\delta(t)$  entre le nombre observé de mort au temps  $t$  et le nombre attendu sous  $H_0$ .
3. Les perdus de vue sont comptabilisés à risque tant qu'ils sont présents...
4. Le test statistique consiste à calculer le carré de la somme des écarts  $\delta(t)$  d'un des deux groupes sur l'ensemble de tous les temps où s'est produit un événement et à le comparer à un seuil  $\Delta(\alpha)$ .

Si le résultat est supérieur au seuil  $\Delta(\alpha)$  on conclut que le risque de mort est différent dans les deux populations.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Exemple : taux de gestation – Vêlage 2 ans

La répartition des échecs de gestation  
par mode de conduite (V2, V3) et par temps de participation

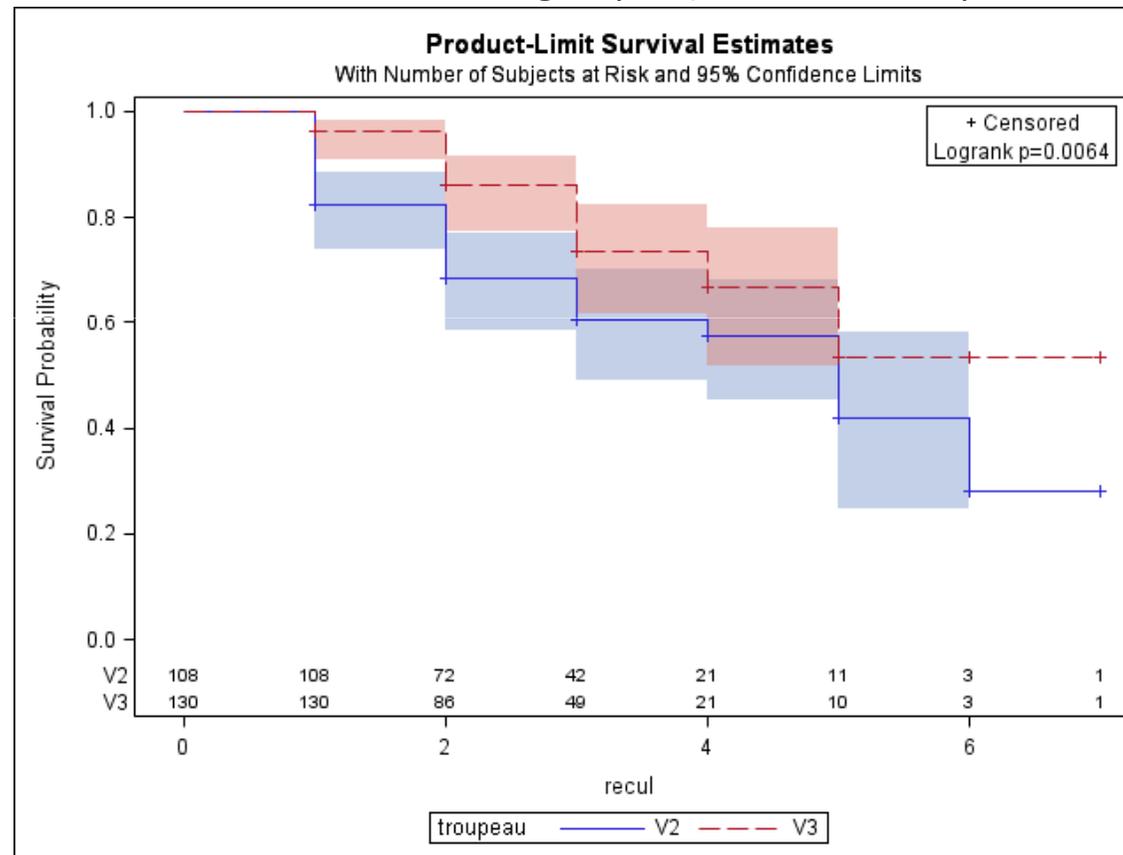
recul	'V2'				'V3'			
	n(V2):à risque	d(V2): échec	'censure'	reste	n(V3):à risque	d(V3): échec	'censure'	reste
1	108	19	17	72	130	5	39	86
2	72	12	18	42	86	9	28	49
3	42	5	16	21	49	7	21	21
4	21	1	9	11	21	2	9	10
5	11	3	5	3	10	2	5	3
6	3	1	1	1	3	0	2	1
7	1	0	1	0	1	0	1	0
		41	67			25	105	

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Exemple : taux de gestation – Vêlage 2 ans

Les courbes de Survie des deux groupes (méthode de Kaplan-Meier)



# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Log-Rank test : la statistique de test

à chaque temps de décès  $t$ ,  $t=1$  à  $k$

	échec	réussite	'à risque'
V2	$d_{V2}(t)$		$r_{V2}(t)$
V3			$r_{V3}(t)$
	$d(t)$	$r(t)-d(t)$	$r(t)$

Conditionnellement aux marges de la table, la statistique 'nombre d'échecs dans le groupe V2'  $d_{V2}(t)$  suit une loi Hypergéométrique de moyenne et de variance :

$$\mu_{V2}(t) = r_{V2}(t) * \frac{d(t)}{r(t)} ; \text{var}(d_{V2}(t)) = \frac{d(t) * (r(t) - d(t)) * r_{V2}(t) * r_{V3}(t)}{r(t)^2 * (r(t) - 1)}$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Log-Rank test : la statistique de test

Au temps t=1

	échec	réussite	'à risque'
V2	19		108
V3			130
	24	214	238

sous l'hypothèse H0 d'égalité des risques d'échecs dans les deux groupes :

$$\text{moy}(dv_2(1) | H_0) = 108 \cdot \frac{24}{238} = 10,9$$

$$\text{var}(dv_2(1) | H_0) = \frac{24 \cdot 214 \cdot 108 \cdot 130}{238^2 \cdot (238 - 1)} = 5,37$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Exemple : taux de gestation – Vêlage 2 ans

(les éléments de calculs)

t=1				t=2				t=3				
	échec	réussite		échec	réussite		échec	réussite		échec	réussite	
V2	19 (10,9)		108	V2	12 (9,6)		72	V2	5 (5,5)		42	
V3	5 (13,1)		130	V3	9 (11,4)		86	V3	7 (6,5)		49	
	24	214	238		21	137	158		12	79	91	

t=4				t=5				t=6				
	échec	réussite		échec	réussite		échec	réussite		échec	réussite	
V2	1 (1,5)		21	V2	3 (2,61)		11	V2	1 (0,5)		3	
V3	2 (1,5)		21	V3	2 (2,4)		10	V3	0 (0,5)		3	
	3	39	42		5	16	21		1	5	6	

*(les effectifs théoriques sous l'hypothèse d'absence d'effet du mode de vêlage)*

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Log-Rank test : la statistique de test

Conditionnellement aux marges les  $k$  variables aléatoires  $\mathbf{d}_{v_2}(\mathbf{t})$  sont indépendantes deux à deux.

On montre alors que la statistique :

$$\chi^2 = \frac{\left( \sum_t d_{v_2}(\mathbf{t}) - \mu_{v_2}(\mathbf{t}) \right)^2}{\sum_t \text{var}(d_{v_2}(\mathbf{t}))}$$

suit une loi de Khi-deux à 1 degré de liberté sous l'hypothèse  $H_0$  d'égalité des deux fonctions de Survie :  $S_{v_2}(t) = S_{v_3}(t)$ .

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### test du Log-Rank : estimation du rapport des risques

On dispose de plusieurs estimateurs du rapport des risques dans les deux groupes (*'Hazard Ratio'*) cf. les compléments.

L'estimateur utilisé est celui de Mantel-Haenszel (1966).

Les estimations du rapport des risques à chaque  $t$  où s'est produit un événement ont également été estimées.

estimation du rapport du risque 'instantané' d'échec à  $t$  :

$$HR(t) = \frac{hv_2(t)}{hv_3(t)}$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Résultats

t	1-Sv2(t)	1-Sv3(t)	hv2(t)	hv3(t)	HR(t)
0	0	0	0	0	--
1	0,176	0,039	0,176	0,038	5,34
2	0,313	0,139	0,167	0,105	1,71
3	0,395	0,262	0,119	0,143	0,81
4	0,424	0,332	0,048	0,095	0,48
5	0,581	0,466	0,273	0,200	1,50
6	0,721	0,466	0,333	0,000	--

Taux d'échec et risques par groupe et rapport de risques HR(t)

	Khi-deux	p
Logrank test	7,44	0,006

$$HR^{MH}(V_2 / V_3) = 2,06$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Log-Rank test : Conditions d'utilisation

#### Conditions d'utilisation du test du Log-Rank (et des tests de la même famille)

Il faut que le risque dans un groupe soit toujours supérieur à celui de l'autre groupe à tous les temps.

Graphiquement il ne faut pas que les courbes se croisent.

Une hypothèse plus forte est que rapport des risques entre les deux groupes reste constant tout au long du temps.

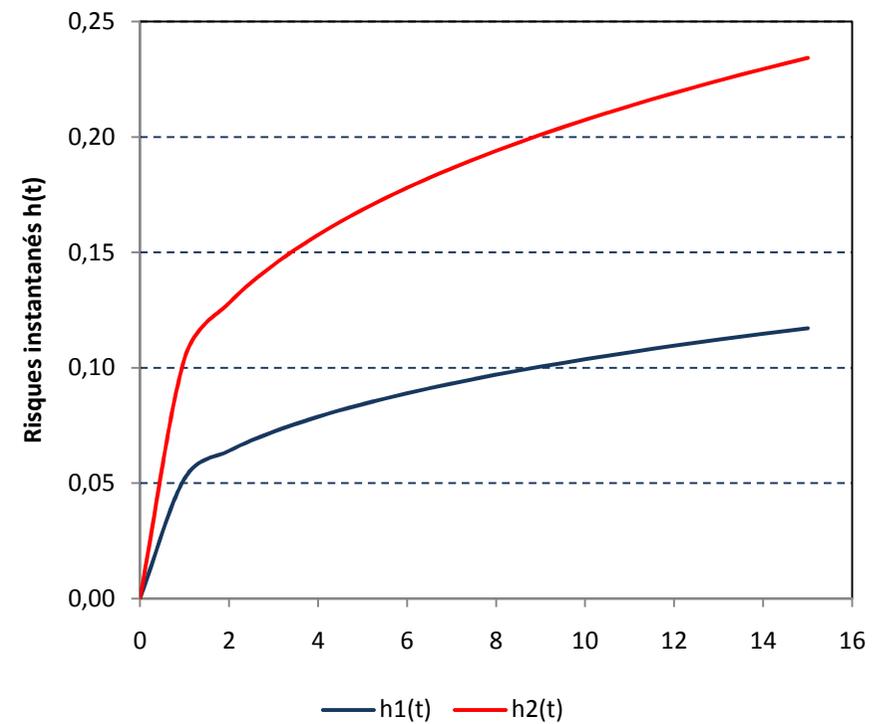
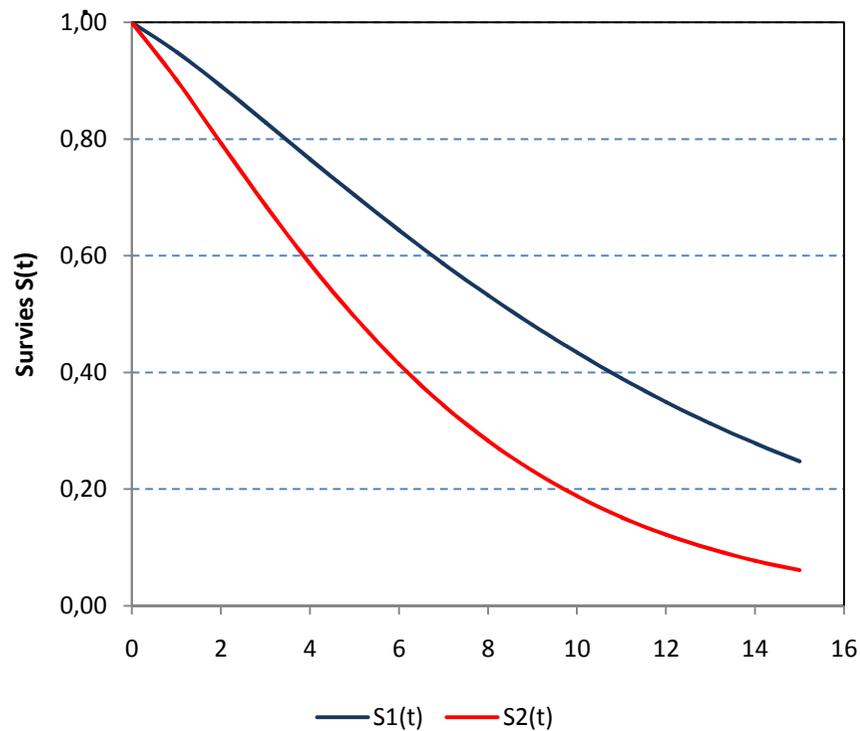
Cette seconde contrainte rend le test du Log-Rank cohérent avec le modèle de Cox.

C'est l'hypothèse de « risques proportionnels » (***Proportional Hazard Models***) nécessaire à l'utilisation du modèle de Cox.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Exemple de modèles à risques proportionnels



$$HR = h_2(t)/h_1(t) = 2 ; \text{ (modèle de Weibull sur } h_1(t)\text{)}$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### **Autres développements autour du test du Log-Rank**

- 1. Comparaisons de plus de deux fonctions de Survie**
- 2. Stratification sur un facteur de confusion**

(voir Compléments)

# LES MÉTHODES SEMI-PARAMETRIQUES

## Le modèle de Cox

## Les méthodes semi-paramétriques

Cas de l'essai « Vêlage 2 ans »

Modélisation du risque d'échec à la gestation  $h(t)$ .

**Groupe V3 (référence)** :  $h(t, V3) = h_0(t)$

**Groupe V2** :  $h(t, V2) = h_0(t) * \eta$      $\eta = \text{constante}$

$\eta$  est le rapport des risques (**'Hazard ratio'**) :  $\eta = \frac{h(t, V2)}{h(t, V3)} = \text{HR}$

**Remarque : ce ratio ne dépend pas du temps.**

Le risque du groupe V2 est proportionnel au risque du groupe V3 tout au long de la carrière des animaux.

Modèle à Risques proportionnels **'Proportional Hazard model'**

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Codage du mode de conduite

Si on définit la variable X indicatrice d'une conduite au vêlage en deux ans :

$$X = \begin{cases} 0 = \text{vêlage 3 ans} \\ 1 = \text{vêlage 2 ans} \end{cases}$$

Le modèle s'écrit :  $h(t, x) = h_0(t) * \exp(x * \beta)$

**Groupe V3 :**  $h(t, 0) = h_0(t) * \exp(0 * \beta) = h_0(t)$

**Groupe V2 :**  $h(t, 1) = h_0(t) * \exp(1 * \beta) = h_0(t) * \exp(\beta)$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Codage du mode de conduite

Le paramètre  $\beta$  mesure l'importance de l'effet de la conduite V2 par rapport à V3.

Si  $\beta = 0$  les risques d'échecs de gestation sont identiques dans les deux groupes

Si  $\beta \neq 0$  les risques sont différents et  $\exp(\beta)$  est une mesure du rapport des risques de V2 par rapport à V3.

$$\text{HR}(V2 / V3) = \exp(\beta)$$

$\beta > 0 \implies \text{HR}(V2 / V3) > 1$  : la conduite V2 est à risque (par rapport à V3)

$\beta < 0 \implies \text{HR}(V2 / V3) < 1$  : la conduite V2 est à 'protectrice'

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Le Modèle de COX (1972)

Le modèle posé par Cox suppose que le Risque est le produit d'une fonction  $h_0(t)$  décrivant le Risque de base au temps  $t$  (*'baseline hazard'*) et d'une fonction exponentielle des facteurs de risque  $X$ .

$$h(t, x) = h_0(t) * \exp(\beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_p * x_p)$$

Exemple (conduite vêlage) :

On veut ajuster l'effet de la conduite 2 ans du mode de reproduction :  
(*'mn'*[ref.], *'s+mn'*, *'i+mn'*)

Le modèle s'écrit :

$$h(t, x) = h_0(t) * \exp(\beta_1 * x_{(v2)} + \beta_2 * x_{(s + mn)} + \beta_3 * x_{(i + mn)})$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Le Modèle de COX

$$h(t, x) = h_0(t) * \exp(\beta_1 * x_{(v2)} + \beta_2 * x_{(s + mn)} + \beta_3 * x_{(i + mn)})$$

#### Définition des variables explicatives

$X_{(v2)}$  = indicatrice de vêlage 2 ans

$X_{(s + mn)}$  = indicatrice du mode de reproduction 's + mn'

$X_{(i + mn)}$  = indicatrice du mode de reproduction 'i + mn'

La quantité  $\exp(\beta_1)$  mesure le rapport des risques d'échec à la gestation pour les animaux conduits en V2 ajusté du mode de reproduction.

Le rapport des risques de deux animaux ayant des profils différents sur les variables explicatives ne dépend pas du temps...

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Le Modèle de COX

#### Profils

Vêlage 3 ans ; 'mn' :  $h(t, \{v3, 'mn'\}) = h_0(t)$

Vêlage 2 ans ; 'mn' :  $h(t, \{v2, 'mn'\}) = h_0(t) * \exp(\beta_1)$

Vêlage 3 ans ; 's + mn' :  $h(t, \{v3, 's + mn'\}) = h_0(t) * \exp(\beta_2)$

Vêlage 2 ans ; 's + mn' :  $h(t, \{v2, 's + mn'\}) = h_0(t) * \exp(\beta_1 + \beta_2)$

...

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Résultats (conduite V2 vs V3)

$$h(t, x) = h_0(t) * \exp(\beta_1 * x_{(v2)})$$

*Ex-æquo* : 'DISCRETE'

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7,4457	1	0,0064
Score	7,4359	1	<b>0,0064 : Log-Rank Test</b>
Wald	7,2137	1	0,0072

### Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Profile Likelihood Confidence Limits
ind_V2 ('DISCRETE')	1	0,7231	0,2692	7,2137	0,0072	<b>2,061</b>	1,224 3,532
ind_V2 ('BRESLOW')	1	0,6403	0,2539	6,3584	0,0117	<b>1,897</b>	1,162 3,160

*ind\_V2 ('DISCRETE')* = Mantel-Haenszel

*ind\_V2 ('BRESLOW')* = estimateur approché

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Résultats (conduite ; mode de reproduction)

$$h(t, x) = h_0(t) * \exp(\beta_1 * X_{(v2)} + \beta_2 * X_{(s + mn)} + \beta_3 * X_{(i + mn)})$$

Modèle sans le poids à la mise à la reproduction (*très peu de données renseignées*)

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Profile Likelihood Confidence Limits	
X(V2)	1	0,71093	0,27399	6,7325	0,0095	2,036	1,198	3,521
X(i + mn)	1	-1,16967	0,78209	2,2367	0,1348	0,310	0,047	1,184
X(s + mn)	1	-0,92695	0,35772	6,7145	0,0096	0,396	0,195	0,797

Un effet de la synchronisation ('s + mn') par rapport à la monte naturelle ajusté de la conduite  
(HR['s + mn' / 'mn'] = 0,396 (p=0,010))

Pas d'effet de l'insémination HR['i + mn' / 'mn'] = 0,310 (p=0,135)

*Manque de puissance ('i + mn': max=10/an ; 's + mn': max=228/an ; 'mn' : max=44/an)*

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Résultats (conduite ; mode de reproduction)

Test de l'interaction du mode de reproduction et de la conduite (V2 vs V3)

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Profile Likelihood Confidence Limits	
X(V2)	1	0,4762	0,471	1,0225	0,3119	1,61	0,642	4,138
X(i + mn)	1	-0,6970	0,826	0,7123	0,3987	0,498	0,072	2,143
X(s + mn)	1	-1,1997	0,494	5,9047	0,0151	0,301	0,114	0,804
X(V2)*X(i + mn)	1	<b>-14,5400</b>	<b>1202,000</b>	0,0001	0,9904	<b>0</b>	<b>.</b>	2,925
X(V2)*X(s + mn)	1	0,4663	0,589	0,6266	0,4286	1,594	0,5	5,095

*Pas d'effet différents du mode de reproduction selon le type de conduite...*

*Test limite pour la modalité 'i + mn' (pas assez de données)*

### Adéquation du modèle de Cox

- **Test de l'hypothèse de risques proportionnels**
  - Graphiquement
  - Par modélisation
- **Influence des données**, analyse des résidus  
Résidus de Cox-Snell, résidus de Martingale, de déviance...  
problème des données censurées  
*(détails cf. Compléments)*

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Adéquation du modèle de Cox

#### Vérification visuelle de l'hypothèse de risques proportionnels

On peut aussi vérifier graphiquement l'hypothèse de risques proportionnels .

$$S(t, x) = \exp[-H(t, x)]$$

$$H(t, x) = H_0(t) \exp(x\beta) \quad \text{Ln}[-\text{Ln}(S(t, x))] = \text{Ln}(H_0(t)) + x\beta$$

$$\text{Vêlage 3 ans :} \quad \text{Ln}[-\text{Ln}(S(t, x))] = \text{Ln}(H_0(t))$$

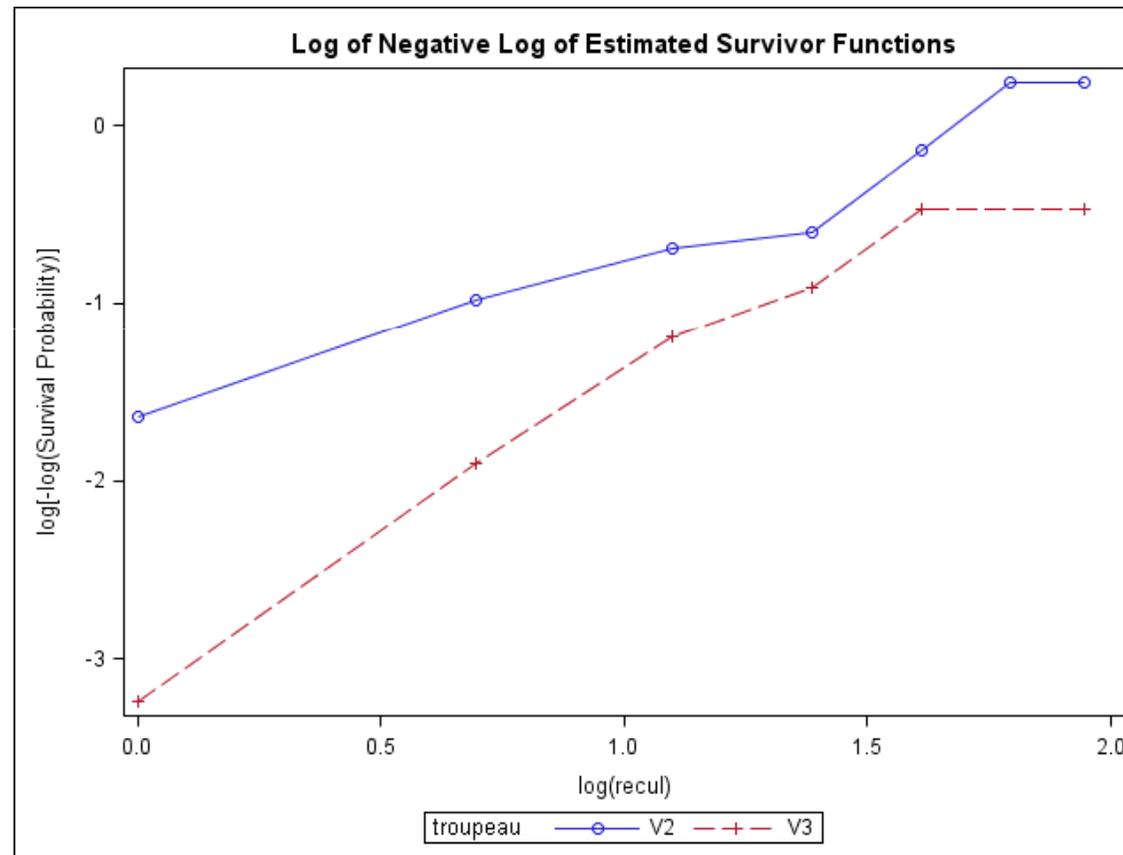
$$\text{Vêlage 2 ans :} \quad \text{Ln}[-\text{Ln}(S(t, x))] = \text{Ln}(H_0(t)) + \beta$$

#### Parallélisme des Survies en échelles LOG(-LOG) des deux groupes

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Adéquation du modèle de Cox : Vêlage 2 ans



### Adéquation du modèle de Cox

#### Test de l'hypothèse de risques proportionnels

$$h(t | V2, X) = HR * h(t | V3, X)$$

HR est indépendant du temps

Cette hypothèse peut être testée à l'aide d'un modèle.

### Adéquation du modèle de Cox

#### Test de l'hypothèse de risques proportionnels par modélisation

L'hypothèse de risques proportionnels sur un facteur de risque X peut se tester en introduisant dans le modèle un terme d'interaction entre ce facteur et le temps.

$$h(t, X) = h_0(t) * \exp(\beta_1 * X + \beta_2 * (X * t))$$

Le test de  $\beta_2$  à zéro teste l'interaction X\*temps .

Si  $\beta_2=0$  l'effet de X sur le risque de survenue de l'événement à t ne dépend pas de t

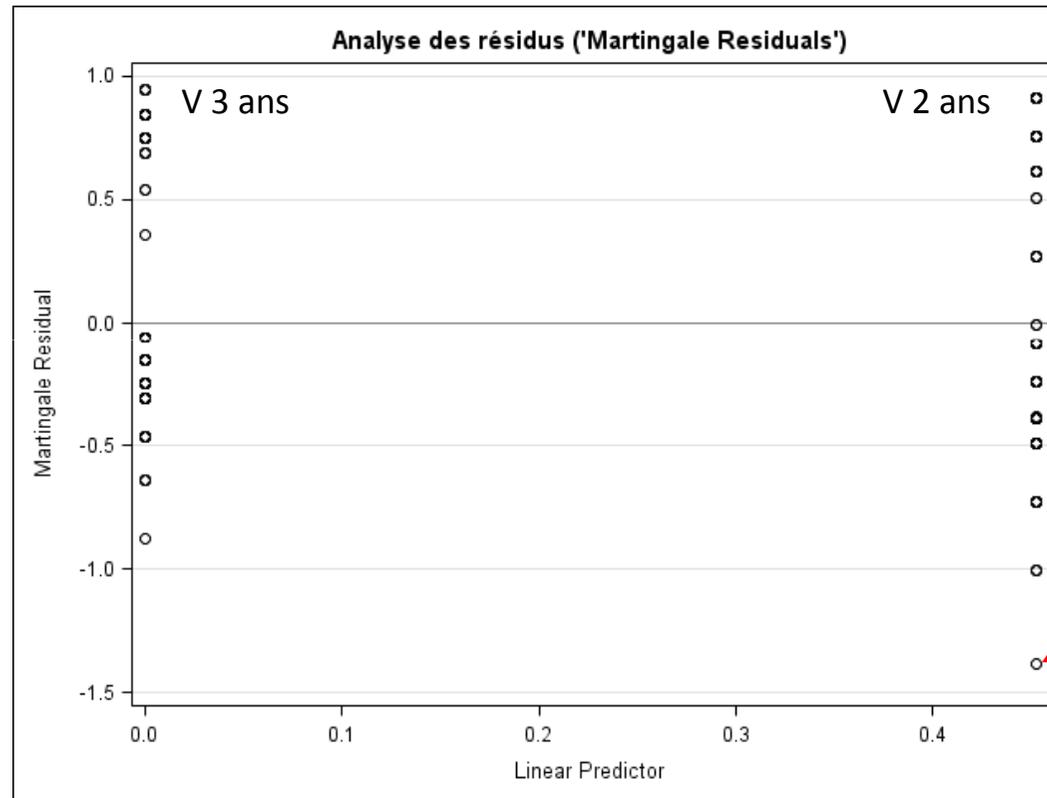
Si  $\beta_2 \neq 0$  l'effet de X s'écrit :  $\beta_1 + \beta_2 * t$ . il dépend du temps t.

Dans l'exemple Vêlage 2 ans le test n'est pas significatif :  $\chi^2_{(1)} = 3,038$  (p=0,081)

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Adéquation du modèle de Cox : analyse des résidus (vêlage 2 ans)



*Modèle sans le mode de reproduction(variable temps-dépendante)*

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

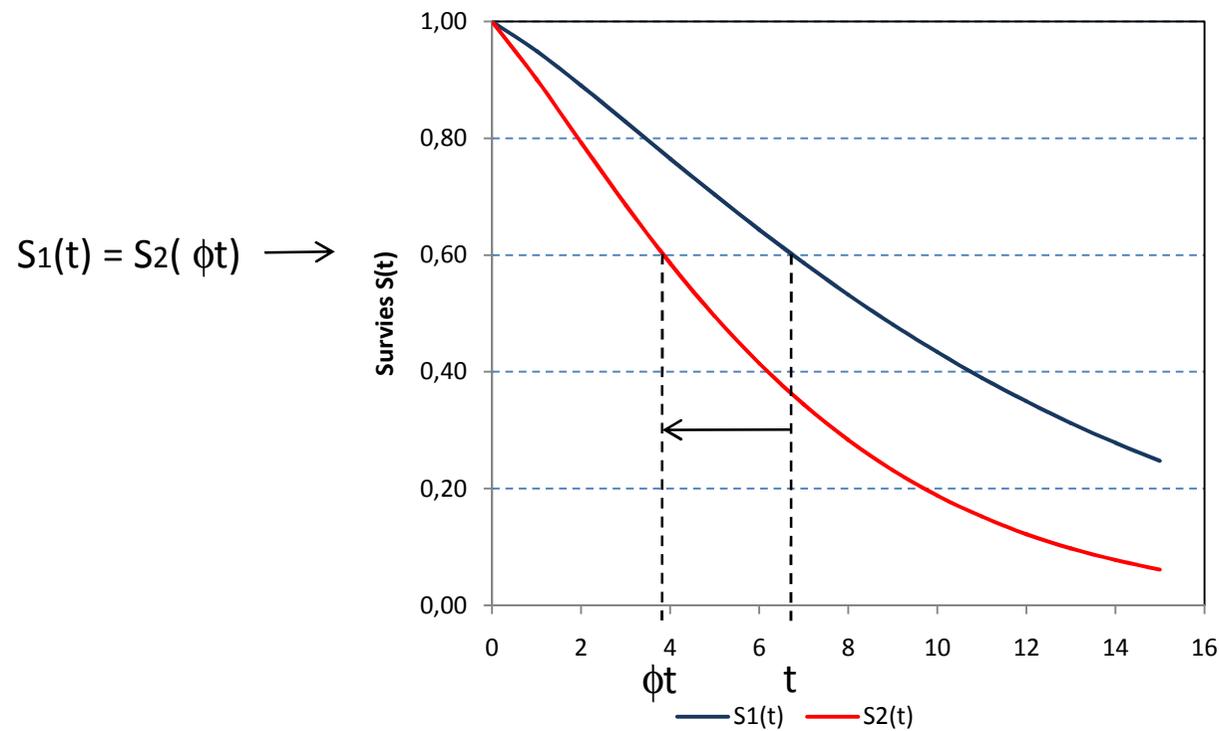
# LES MÉTHODES PARAMETRIQUES

## Les modèles à temps accélérés

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Les modèles à temps accélérés



$S_1(t)$  : courbe de référence ( $\phi < 1$ )

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Les modèles à temps accélérés

$\phi < 1$  : A même probabilité de survie, l'événement se produit plus tôt dans le groupe 2 que dans le groupe 1.

« Accélération » du temps d'apparition de l'événement

$$S_1(t) = S_2(\phi * t)$$

Le facteur d'accélération  $\phi$  représente l'exposition à des facteurs de risque X

$$\phi = \exp(X * \beta)$$

*Remarque* :  $\phi < 1 \Rightarrow \beta < 0$

Différentes modélisations de la fonction de risque instantané  $h(t)$ ...

### Le modèle de Weibull

Weibull( $\lambda ; \gamma$ )     $\lambda$  = paramètre d'échelle

$\gamma$  = paramètre de forme

$$S(t) = \exp(-\lambda * t^\gamma) \quad ; \quad h(t) = \lambda * \gamma * t^{(\gamma-1)}$$

**$\gamma = 1 \rightarrow$  modèle exponentiel**

Particularité : la fonction de risque est monotone.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Fonctions de Risque $h(t)$ pour différentes lois de Weibull

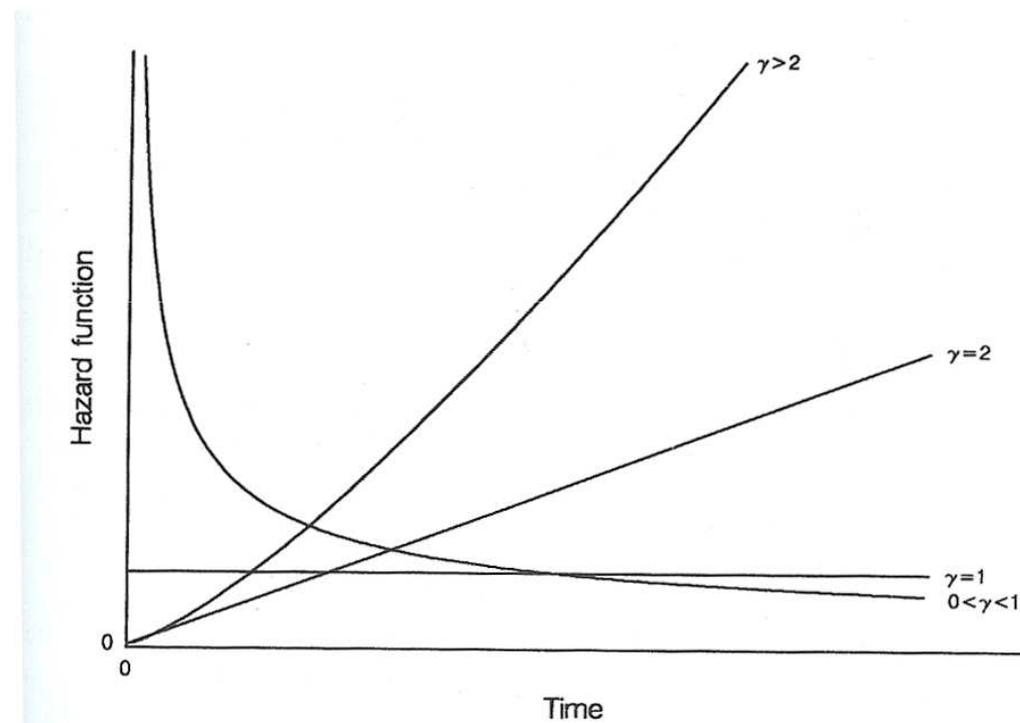


Figure 4.3 The form of the hazard function  $h(t) = \lambda\gamma t^{\gamma-1}$  for different values of  $\gamma$ .

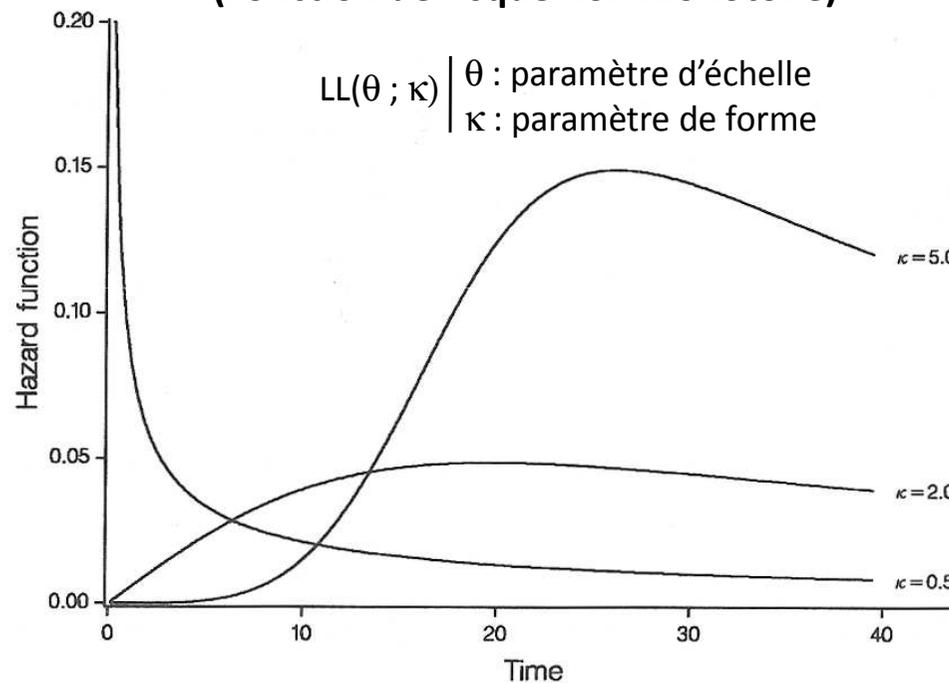
(Collett, D. 1995)

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Fonctions de Risque $h(t)$ pour la loi Log-logistique

(Fonction de risque non monotone)



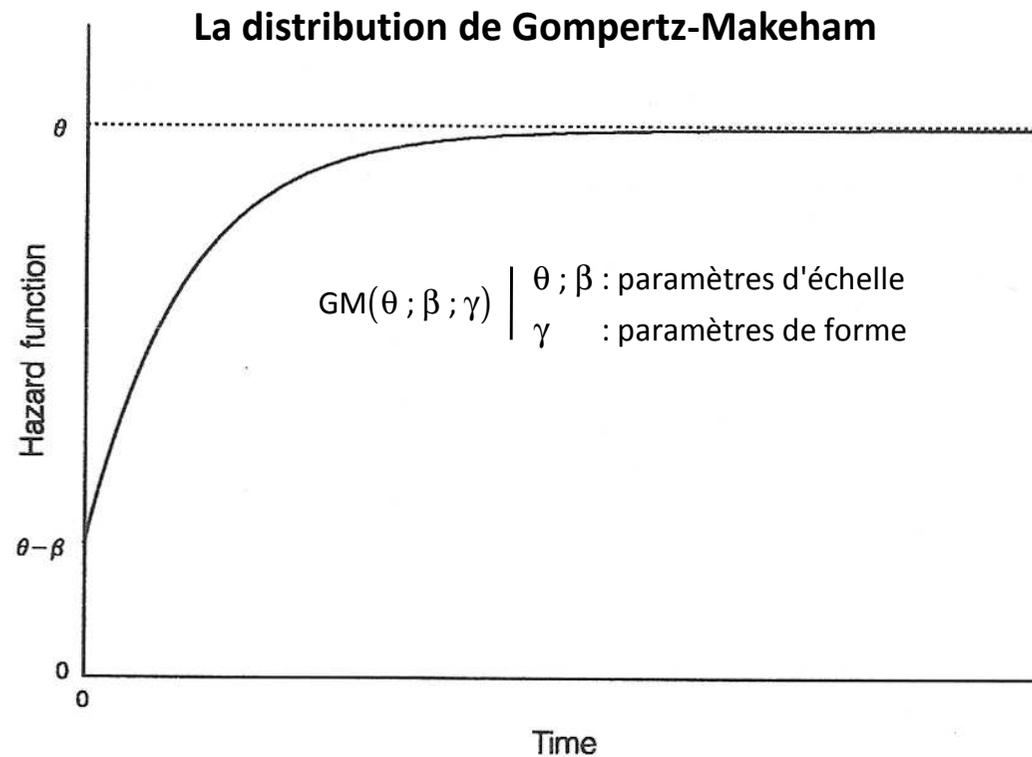
**Figure 6.1** Hazard functions for a log-logistic distribution with a median of 20 and  $\kappa = 0.5, 2.0$  and  $5.0$ .

(Collett, D. 1995)

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Fonctions de Risque $h(t)$ avec asymptote



**Figure 6.2** An asymptotic hazard function, where  $h(t) = \theta - \beta e^{-\gamma t}$ .

(Collett, D. 1995)

### Le modèle à temps accéléré dans le cas d'une distribution de Weibull

Un modèle Log-linéaire sur le temps :

Pour un individu décrit sur un profil X de variables (facteurs de risque)

t = temps au bout duquel apparaît un événement donné (la 'mort')

$$\text{Log}_e(t | x) = \mu + x * \alpha + \sigma * \varepsilon$$

Où :

- $\mu$  et  $\sigma$  sont des paramètres caractéristiques de la loi de distribution des temps dans la population de base (X=0) ('baseline')
- $\alpha$  est un ensemble de paramètres mesurant l'influence des variables X sur le temps t de survenue de l'événement étudié.

Exemple des veaux...

## Le modèle à temps accéléré dans le cas d'une distribution de Weibull

### Exemple des veaux

Etude du risque de pneumonie suivant deux types de logements des veaux.

$X = (0 \text{ bâtiment } \mathbf{0} \text{ (référence)} / 1 \text{ bâtiment } \mathbf{1})$

Bâtiment 0 :	$\text{Log}_e(t 0) = \mu$
Bâtiment 1 :	$\text{Log}_e(t x) = \mu + \alpha$

Si le bâtiment **1** est plus à risque que le bâtiment **2** l'événement se produit avant dans ce bâtiment  $\implies \alpha < 0$ .

Si les deux bâtiment ont les mêmes risques  $\alpha = 0$

Si le bâtiment **1** est moins à risque que le bâtiment **2** l'événement se produit après dans ce bâtiment  $\implies \alpha > 0$ .

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Une particularité du modèle de Weibull

**Le modèle de Weibull est un modèle à risque proportionnels.**

Le risque dans la population exposée s'exprime à partir des paramètres  $\beta$  caractérisant les effets des facteurs de risque  $X$  sur la fonction de risque instantané  $h(t, X)$  et du paramètre de forme  $\gamma$  de la loi de Weibull posée sur la fonction de risque instantané  $h_0(t)$  la population de référence ( $X=0$ ).

$$h(t, x) = (\exp(x * \beta))^{\gamma} * h_0(t) \quad \text{Rapport de risques : } HR(x) = \exp[\beta * \gamma]$$

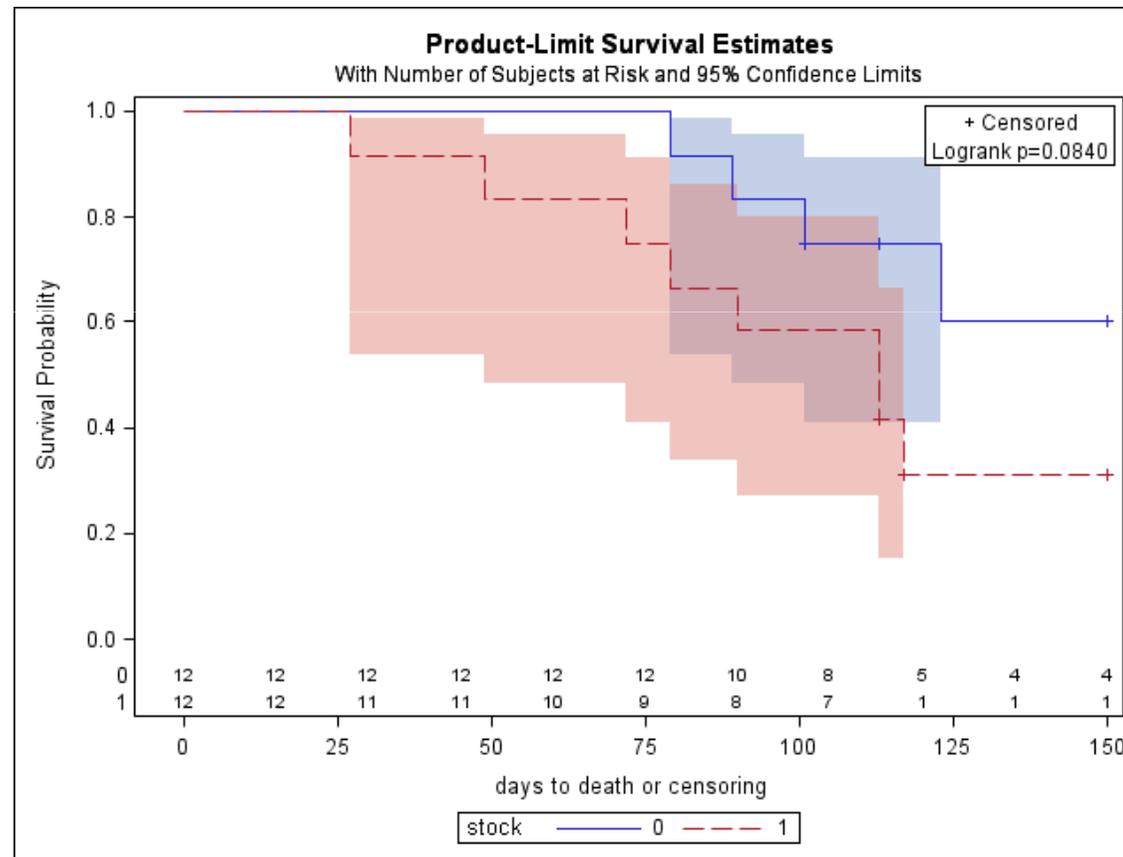
Relation entre les paramètres  $\alpha$  du modèle à Temps accélérés et les paramètres  $\beta$  du modèle à risques proportionnels.

$$\gamma = \frac{1}{\sigma} \quad ; \quad \beta = \frac{-\alpha}{\sigma}$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Exemple des pneumonies de veaux



# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Exemple des pneumonies de veaux

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5,2338	0,213	4,8162	5,6513	603,66	<,0001
$\hat{\alpha}$ : stock	1	-0,4300	0,247	-0,9141	0,0542	3,03	0,0817
$\hat{\sigma}$ : Scale	1	0,3819	0,0974	0,2316	0,6295		
$\hat{\gamma}$ : Weibull Shape	1	2,6188	0,668	1,5885	4,3174		

$$\text{HR}(\text{Stock}=1/\text{Stock}=0) = \exp(-(-0,4300) / 0,3819) = \exp(0,4300 * 2,6188) = 3,08$$

Pour mémoire :

HR[Cox] = 2,85 (p=0,097)

HR[Log-Rank Test] = 2,70 (p=0,084) (*formule approchée cf. Compléments*)

### Adéquation du modèle de Weibull

La fonction de Survie dans le cadre du modèle de Weibull s'écrit :

$$S(t) = \exp(-\lambda * t^\gamma)$$

On peut en déduire une expression de la Survie qui dépend linéairement du logarithme du temps :

$$\text{Log}_e[-\text{Log}_e(S(t))] = \text{Log}_e(\lambda) + \gamma * \text{Log}_e(t)$$

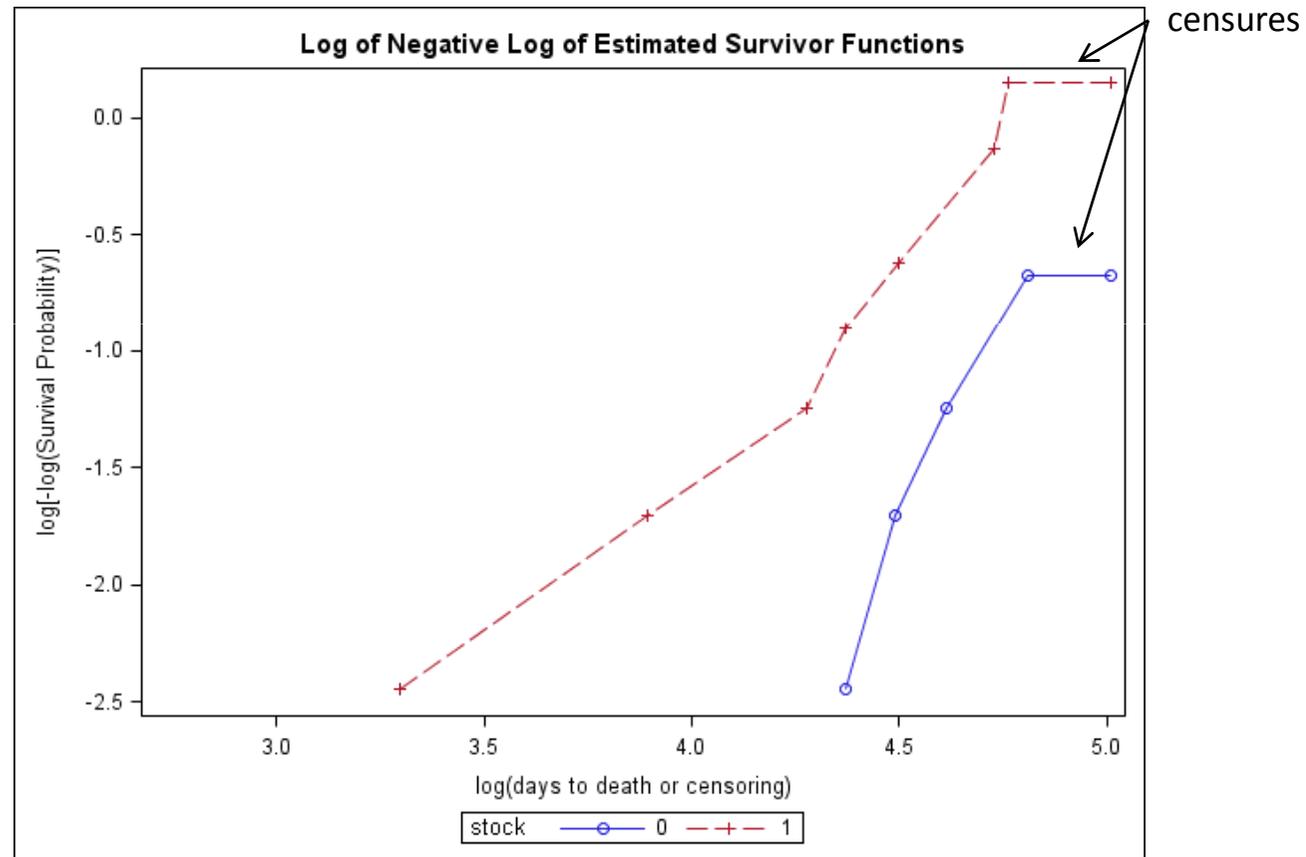
Si le modèle de Weibull s'applique correctement aux données les variations de la survie en échelle LOG(-LOG) sont proportionnelles au logarithmes des temps d'observations.

Cette expression, appliquée à chacune des populations de veaux, permet de vérifier graphiquement l'adéquation des ajustements par une Weibull.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

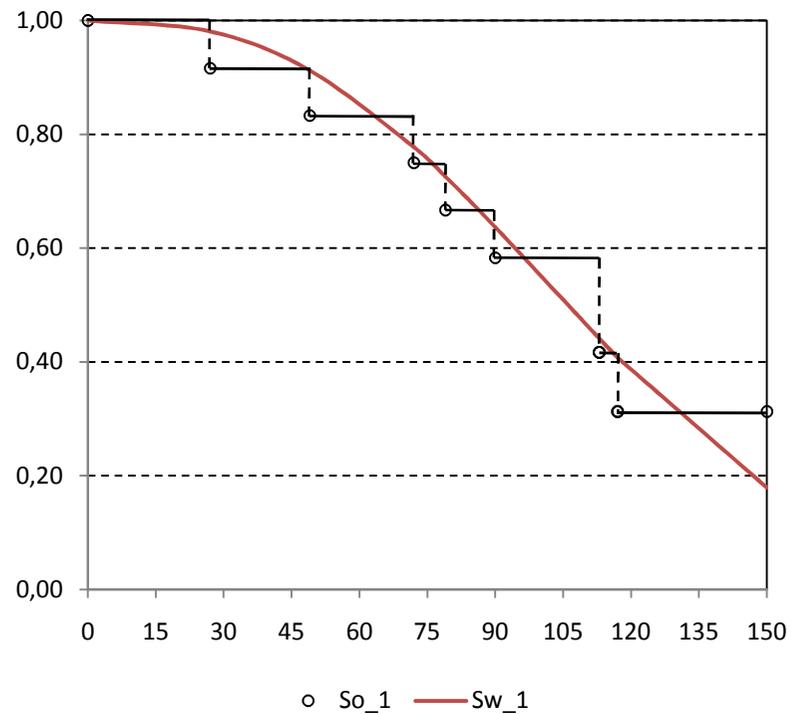
### Adéquation du modèle de Weibull (ex. des veaux)



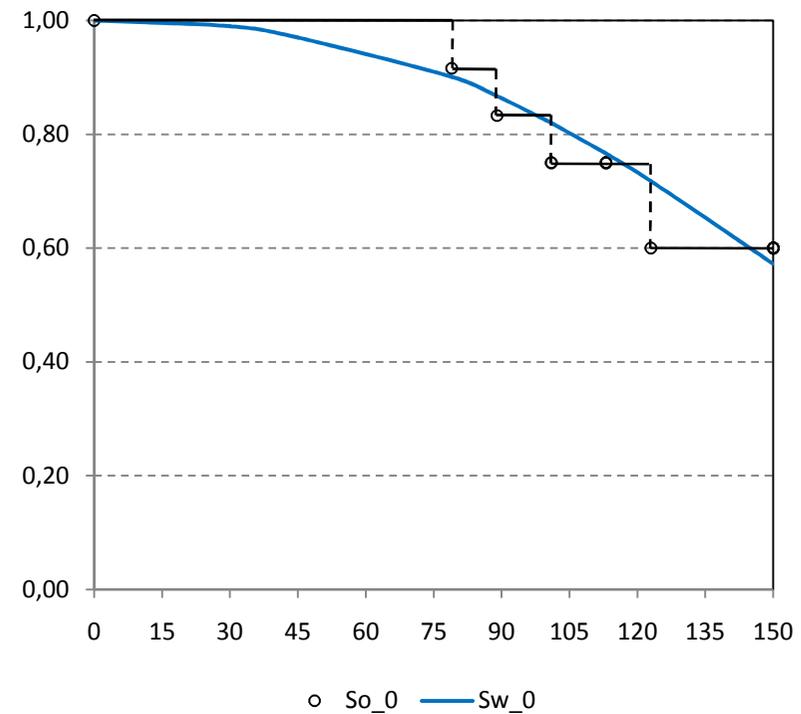
# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Adéquation du modèle de Weibull



Stock 1



Stock 0

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### **ON NE VERRA PAS ...**

**L'introduction de variables temps-dépendantes dans le modèle de Cox.**

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Éléments de bibliographie (articles)

- Bernstein, L., Anderson, J., Pike, M.C. (1981). Estimation of the proportional hazard in two-treatment-group. *Biometrics*, **37**, 513-519.
- Cox, D.R., Snell, E.J. (1968). A general definition of residuals. *J. R. Stat. Soc. B*, **30**, 248-275.
- Crowley, J. and Hu, M. (1977), "Covariance Analysis of Heart Transplant Survival Data," *Journal of the American Statistical Association*, **72**, 27–36.
- Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, **52**, 203-223.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemother. Rep.*, **50**, 163-170.
- Moreau, T., Le Minor, M., Myquel, P., Lellouch, J. (1985). Estimation of the hazards ratio in two grouped samples. *Biometrics*, **41**, 245-252.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Éléments de bibliographie (articles)

- Peto, R., Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J. R. Stat. Soc. A*, **135**, 185-207.
- Prentice, R.L. (1978). Linear rank tests with right censored data. *Biometrika*, **65**, 167-179.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Éléments de bibliographie (livres)

- Collett D. (1994) : *Modelling survival data in medical research* - Chapman & Hall
- Cox D.R., Oakes D. (1984) : *Analysis of Survival data* - Chapman & Hall
- Dohoo J., Martin W., Stryhn H. (2003) : *Veterinary Epidemiologic Research* National Library of Canada
- Hill C., Com-Nougué C., Kramar A., Moreau T., O'Quigley J., Senoussi R., Chastang C. : *Analyse statistique des données de survie* (1990) INSERM Medecine-Sciences - Flammarion
- Hougaard P. (2000) : *Analysis of Multivariate Survival Data* – Springer
- Kalbfleisch, J. D. and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data 2d ed*, New York: John Wiley & Sons.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

# Compléments

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

- Compléments sur le test du Log-Rank : le test
- Compléments sur le test du Log-Rank : Estimation du rapport des risques
- Compléments sur le test du Log-Rank : Généralisation à d'autres tests
- Compléments sur le test du Log-Rank : cas de plus de deux populations
- Test du Log-Rank stratifié
- La Vraisemblance partielle de Cox
- L'adéquation du modèle de Cox
- La modélisation de différentes transformations du Risque (Modèle de Cox)
- Le modèle de Weibull est un modèle à Risques Proportionnels

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

### Compléments sur le test du Log-Rank : le test

à chaque temps de décès  $t$ ,  $t=1$  à  $k$

	échec	réussite	'à risque'
V2	$dv_2(t)$		$rv_2(t)$
V3	$dv_3(t)$		$rv_3(t)$
	$d(t)$	$r(t)-d(t)$	$r(t)$

On désigne par  $O_1$ ,  $O_2$  respectivement le nombre d'évènements observés sur l'ensemble des  $k$  tables dans le premier groupe ('V2'), dans le second groupe ('V3').

$$O_1 = \sum_{t=1}^{t=k} dv_2(t) \quad ; \quad O_2 = \sum_{t=1}^{t=k} dv_3(t)$$

De façon analogue on désigne par  $E_1$ ,  $E_2$  le nombre d'évènements attendus sous l'hypothèse  $H_0$  de survie identique dans les deux groupes.

$$E_1 = \sum_{t=1}^{t=k} rv_2(t) * \frac{d(t)}{r(t)} \quad ; \quad E_2 = \sum_{t=1}^{t=k} rv_3(t) * \frac{d(t)}{r(t)}$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Compléments sur le test du Log-Rank : le test

Le test du Log-Rank est une forme particulière du test d'ajustement de Mantel-Haenszel.

Le temps est l'analogie du facteur de confusion sur lequel est ajustée la statistique de test dans le test de Mantel-Haenszel.

Il existe **une forme simplifiée du test** du Log-Rank dans le cas de deux groupes.

Si on appelle  $O_1$ ,  $O_2$ ,  $E_1$ ,  $E_2$ , respectivement, le nombre d'évènements observés dans les groupes 1 et 2, le nombre d'évènements attendus sous l'hypothèse de survie identique dans les deux groupes, alors la statistique :

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

est toujours inférieure à la statistique du Log-Rank et fournit un test approché conservateur.

### Compléments sur le test du Log-Rank : estimation du rapport des risques

On dispose de plusieurs estimateur du rapport des risques dans les deux groupes (*'Hazard Ratio'*).

#### 1. L'estimateur de Mantel-Haenszel (1966)

$$HR^{MH}_{(v2 / v3)} = \frac{\sum_{t=1}^k d_{v2}(t) * (r_{v3}(t) - d_{v3}(t)) / r(t)}{\sum_{t=1}^k d_{v3}(t) * (r_{v2}(t) - d_{v2}(t)) / r(t)}$$

#### 2. L'estimateur de Tarone (1981)

$$HR^T_{(v2 / v3)} = OR_{MH} = \frac{\sum_{t=1}^k d_{v2}(t) * r_{v3}(t) / r(t)}{\sum_{t=1}^k d_{v3}(t) * r_{v2}(t) / r(t)}$$

### Compléments sur le test du Log-Rank : estimation du rapport des risques

#### 3. Un estimateur approché

Cet estimateur est basé sur le nombre d'évènements observés ( $O_1$ ,  $O_2$ ) et attendus ( $E_1$ ,  $E_2$ ) dans les deux groupes.

$$HR^{ap}_{(v_2 / v_3)} = \frac{O_1/E_1}{O_2/E_2}$$

Cet estimateur doit être utilisé avec précaution car c'est un estimateur biaisé dont le biais est d'autant plus important que le vrai rapport des risques est élevé et que les effectifs sont grands (Bernstein, 1981).

Les estimateurs de Mantel-haenszel et de Tarone sont asymptotiquement sans biais quelle que soit la vraie valeur des risques instantanés. En présence d'ex-aequo l'estimateur de Mantel-Haenszel présente un biais négligeable à condition que ceux-ci soient en nombre limité. C'est l'estimateur recommandé (Moreau, 1985).

### Compléments sur le test du Log-Rank : généralisation

D'autres statistiques peuvent être utilisées pour comparer des fonctions de Survie.

On peut montrer que toutes les statistiques s'écrivent suivant la forme générale suivante :

$$\chi^2 = \frac{\left( \sum_t \omega(t) (d_{v2}(t) - \mu_{v2}(t)) \right)^2}{\sum_t \omega(t)^2 * \text{var}(d_{v2}(t))}$$

où le poids  $\omega(t)$  prend des valeurs différentes suivant les statistiques de test.

**1.  $\omega(t) = 1$  : Logrank-test**

C'est la pondération la plus simple. Elle attribue le même poids à tous les évènements quel que soit le moment où ils se produisent.

**2.  $\omega(t) = r(t)$  : Gehan (Wilcoxon)**

Le poids est égal au nombre d'individus à risque à  $t$ . Ce système de pondération, proposé par Gehan (1965) privilégie les évènements précoces.

### Compléments sur le test du Log-Rank : généralisation

#### 3) $\omega(t) = s^*(t)$ : Peto-Prentice

Les poids sont les valeurs de la survie sous l'hypothèse de survie identique dans les deux groupes. Ce système de pondération suggérés par Peto et Prentice (Prentice, 1978) privilégie les évènements précoces comme la pondération de Gehan.

D'autres pondérations sont proposées (cf. la documentation SAS ...).

#### Lequel choisir ?

Le test du Log-Rank est le plus puissant en absence de censure et en présence de censure si on peut faire l'hypothèse de l'indépendance entre les processus de censure et de survie.

Les tests de Gehan et Peto-Prentice sont plus aptes à déceler des différences en cas de morts précoces. La statistique de Gehan dépend davantage de la distribution des censures que celle de Peto-Prentice. Il est préférable de ne pas l'utiliser.

L'usage semble privilégier le test du Log-Rank notamment en raison de la cohérence de ses résultats avec certains résultats du modèle de Cox.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Compléments sur le test du Log-Rank : comparaison de plus de deux populations

La méthode se généralise à la comparaison globale de  $q$  fonctions de survie ( $q > 2$ ).

La statistique est formée à partir du vecteur  $\underline{U}$  composé de la somme des écarts entre le nombre d'évènements observés et attendus pour  $(q-1)$  groupes parmi les  $q$ .

$$\underline{U} = \begin{bmatrix} O_1 - E_1 \\ \dots \\ O_j - E_j \\ \dots \\ O_{(q-1)} - E_{(q-1)} \end{bmatrix} \quad O_j = \sum_t dv_j(t) \quad ; \quad E_j = \sum_t rv_j(t) * \frac{d(t)}{r(t)}$$

où :

$dv_j(t)$  désigne le nombre d'évènements observés dans le groupe  $j$  au temps  $t$ .

$rv_j(t)$  désigne le nombre d'individus à risque dans le groupe  $j$  au temps  $t$ .

$d(t)$  désigne le nombre d'évènements observés sur l'ensemble des  $K$  groupes au temps  $t$ .

$r(t)$  désigne le nombre d'individus à risque sur l'ensemble des  $q$  groupes au temps  $t$ .

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Compléments sur le test du Log-Rank : comparaison de plus de deux populations

Ce vecteur  $\mathbf{U}$  a une matrice de variance-covariance  $\mathbf{W}$  qui généralise la variance de la loi Hypergéométrique du cas à deux populations.

Sous l'hypothèse  $H_0$  d'égalité des  $q$  fonctions de Survie  $\mathbf{W}$  s'écrit :

$$\text{var}(d_{vj}(t)) = \frac{d(t) * (r(t) - d(t)) * r_{vj}(t) * (r(t) - r_{vj}(t))}{r(t)^2 * (r(t) - 1)} \quad j = 1, \dots (q-1)$$

$$\text{cov}(d_{vj}(t); d_{vj'}(t)) = -\frac{d(t) * (r(t) - d(t)) * r_{vj}(t) * r_{vj'}(t)}{r(t)^2 * (r(t) - 1)} \quad j, j' = 1, \dots (q-1) ; j' \neq j$$

$$\text{cov}(d_{vj}(t); d_{vj'}(t')) = 0 \quad t \neq t'$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Compléments sur le test du Log-Rank : comparaison de plus de deux populations

On montre que sous  $H_0$  « égalité des  $q$  fonctions de Survie » le vecteur  $\underline{U}$  a une distribution asymptotique Normale à  $q-1$  dimension et la statistique

$$\chi^2 = \underline{U}' * W^{-1} * \underline{U}$$

suit une distribution de Khi-deux à  $(q-1)$  degrés de libertés.

La statistique  $\chi^2$  est alors comparée au percentile  $(1-\alpha/2)$  d'un Khi-deux à  $(q-1)$  degrés de libertés.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Compléments sur le test du Log-Rank : stratification

Il est possible de prendre en compte une structuration des données en  $S$  strates dans la comparaison de courbes de survie.

Les effectifs attendus sous l'hypothèse nulle d'égalité des fonctions de survie sont calculés pour chaque strate  $j$ .

$$E_1(j) = \sum_{t=1}^{t=k} r_{v_2}(t) * \frac{d(t)}{r(t)} \quad ; \quad E_2(j) = \sum_{t=1}^{t=k} r_{v_3}(t) * \frac{d(t)}{r(t)} \quad \text{pour chaque strate } j, j=1 \text{ à } S$$

*(cas de deux groupes)*

L'écart  $d(j)$  entre effectifs observé et attendu est calculé par strate pour l'un des deux groupes dans le cas de deux groupes à comparer, pour  $(q-1)$  groupes dans le cas de la comparaison de  $q$  groupes.

$$d(j) = O_1(j) - E_1(j) \quad \text{où} \quad O_1(j) = \sum_{t=1}^{t=k} dv_2(t)$$

### Compléments sur le test du Log-Rank : stratification

La statistique globale sur l'ensemble des  $S$  strates est le carré de la somme  $\mathbf{d}$  sur les  $S$  strates des écarts  $d(j)$  par strate rapportée à la variance de  $\mathbf{d}$ .

$$\chi^2 = \frac{(\mathbf{d})^2}{\text{var}(\mathbf{d})} \quad \text{où} \quad \mathbf{d} = \sum_{j=1}^S d(j)$$

Comme les  $S$  écarts  $d(j)$  sont indépendants la variance de  $\mathbf{d}$  est la somme des variances des écarts par strate.

$$\text{var}(\mathbf{d}) = \sum_{j=1}^S \text{var}(d(j))$$

Sous l'hypothèse  $H_0$  d'égalité des fonctions de survie la statistique  $\chi^2$  suit une distribution de Khi-deux à 1 degré de liberté.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### La vraisemblance partielle du modèle de COX

Le Risque de base  $h_0(t)$  est un paramètre de nuisance pour l'estimation des paramètres  $\beta$ . Cox s'en débarrasse en raisonnant conditionnellement aux 'morts'.

Les données sont classées suivant le temps de participation  $t$  dans l'ordre de survenue des évènements, morts 'M' et censures 'C'.

On a ainsi la série d'évènements suivante :

$$[E] = \{C_1, M_1, C_2, M_2 \dots, C_i, M_i \dots, C_k, M_k, C_{k+1}\}$$

Où  $M_i$  désigne l'évènement « Mort en  $t_i$  » et  $C_i$  est l'évènement associé à l'ensemble des censures observées entre  $t_{i-1}$  et  $t_i$ . (pour  $C_1$ , entre  $t_0=0$  et  $t_1$  et pour  $C_{k+1}$ , les censures observées après  $t_k$ ).

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### La vraisemblance partielle du modèle de COX

Si on raisonne conditionnellement aux temps d'observation des morts

La vraisemblance associée aux évènements E s'écrit :

$$Vr([\underline{E}]) = Vr(C_1) * Vr(M_1 | C_1) * Vr(C_2 | C_1, M_1) \dots * Vr(C_{k+1} | C_1 \dots C_2, M_1 \dots M_k)$$

$$Vr([\underline{E}]) = \prod_{i=1}^k Pr[M_i | C^{(i)}, M^{(i-1)}] * \prod_{i=1}^{k+1} Pr[C_i | C^{(i-1)}, M^{(i-1)}]$$

Où  $C_i^{(i)}$  désigne la suite d'évènements  $C_1, C_2 \dots C_i$

$M_i^{(i)}$  désigne ma suite des évènements  $M_1, M_2 \dots M_i$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### La vraisemblance partielle du modèle de COX

Si on admet que le second terme de la vraisemblance contient peu d'information sur les paramètres  $\beta$  (ce qui revient à supposer que le processus des censures est indépendant de celui des mortalités) seul le premier terme est porteur d'informations sur les paramètres  $\beta$ .

C'est cette « vraisemblance partielle » que maximise Cox (encore appelée « Vraisemblance de Cox »)

$$V^*r([\underline{E}]) = \prod_{i=1}^k \Pr[M_i | C^{(i)}, M^{(i-1)}]$$

Cette vraisemblance revient à exprimer la probabilité pour un individu 'vivant' juste avant  $t$  par rapport à l'ensemble des  $r(t_i)$  individus à risque à  $t_i$  (Probabilité conditionnelle à la population à risque à  $t_i$ ).

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### La vraisemblance partielle du modèle de COX

*cas sans ex-æquos (un 'mort' et un seul à chaque temps  $t_i$ )*

$R(t_i)$  individus à risque à  $t_i$  : La probabilité que l'individu  $i$  de profils  $x_i$  soit mort à  $t_i$  s'écrit :

$$\Pr(\text{individu mort en } t_i | R(t_i)) = \frac{h(t_i, x_i)}{\sum_{j \in R(t_i)} h(t_i, x_j)} = \frac{\exp(x_i \beta)}{\sum_{j \in R(t_i)} \exp(x_j \beta)}$$

Cette probabilité est indépendante de  $h_0(t)$ .

La vraisemblance de Cox se calcule alors comme le produit de la contribution de tous les individus à risque suivant leurs statuts (morts, vivants) à tous les temps  $t_1$  à  $t_k$ .

## La vraisemblance partielle du modèle de COX

### *cas avec ex-æquo (deux 'morts' ou plus observés simultanément)*

Il faut tenir compte de toute la combinatoire des ordres possibles .

Ainsi si deux individus sont morts en même temps  $e_1, e_2$  il y a deux ordres possibles  $\{e_1, e_2\}$  et  $\{e_2, e_1\}$ .

Avec trois individus  $e_1, e_2, e_3$  il y a 6 ordres possibles :

$\{e_1, e_2, e_3\}, \{e_1, e_3, e_2\}, \{e_2, e_1, e_3\}, \{e_2, e_3, e_1\}, \{e_3, e_1, e_2\}, \{e_3, e_2, e_1\}$

Avec  $v$  individus il y a  $v!$  ordres possibles...

Dans le cas de 2 ex-æquo,  $e_1, e_2$  leur contribution à la vraisemblance de Cox s'écrit:

$$\frac{\exp(x_1\beta)}{\sum_{j \in R(t_1)} \exp(x_j\beta)} * \frac{\exp(x_2\beta)}{\sum_{j \in R(t_1)-e_1} \exp(x_j\beta)} + \frac{\exp(x_2\beta)}{\sum_{j \in R(t_1)} \exp(x_j\beta)} * \frac{\exp(x_1\beta)}{\sum_{j \in R(t_1)-e_2} \exp(x_j\beta)}$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Adéquation du modèle de Cox, analyse des résidus

Cox et Snell (1968) ont proposé une expression des résidus largement utilisée en définissant le résidu de l'individu  $i$  comme étant l'estimation du risque cumulé pour cet individu jusqu'au moment de l'apparition de l'événement étudié (résidus de Cox-Snell).

$$r_{ci} = \hat{H}(t_i) \quad \text{où} \quad H(t_i) = H_0(t_i) * \exp(x_i * \beta)$$

où  $H_0(t_i)$  désigne le risque cumulé de base au temps  $t_i$  (à niveau nul des covariables  $X$ ) et  $x_i$  représente le profil des valeurs des covariables caractérisant l'individu  $i$ .

On démontre que quelle que soit la distribution de la fonction de Survie la fonction de risque cumulée suit une distribution exponentielle de paramètre  $\lambda=1$ .

On considère alors que si le modèle ajuste bien les données la distribution des estimations des risque cumulées  $\hat{H}(t_i)$  suit également une distribution exponentielle de paramètre  $\lambda=1$ .

Cela permet d'avoir un cadre distributionnel utile à l'examen des résidus.

La distribution des résidus de Cox-Snell est positive fortement asymétrique autour de 1.

### Adéquation du modèle de Cox, analyse des résidus

#### Les résidus de Cox-Snell modifiés

Les observations censurées posent un problème avec cette définition car l'estimation du risque cumulé est biaisée inférieurement.

$$t_i^* \Rightarrow \hat{H}(t_i^*) < H(t_i) \quad (\text{où } t_i^* \text{ désigne le temps censuré})$$

Des résidus de Cox-Snell modifiés ont été proposés (Crowley-Hu, 1977) qui permettent de proposer des résidus pour les observations censurées.

$$r'_{ci} = \left\{ \begin{array}{ll} r_{ci} & \text{si l'observation n'est pas censurée} \\ r_{ci} + 0,693 & \text{si l'observation est censurée} \end{array} \right\}$$

Le résidu des observations censurée est calculé en rajoutant au risque cumulé jusqu'à la date de censure un terme qui correspond au temps médian de survie d'une distribution exponentielle de paramètre  $\lambda=1$  ( $S(0,693) = \exp(-0,693)=1/2$ ).

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Adéquation du modèle de Cox, analyse des résidus

#### Les résidus de Martingale

A partir des résidus de Cox –Snell des résidus de moyenne nulle ont été proposés, les résidus de Martingale (*'martingale residuals'*).

$$r_{Mi} = \delta_i - r_{Ci} \quad \text{où } \delta_i = \begin{cases} 0 & \text{si observation censurée} \\ 1 & \text{si observation non censurée} \end{cases}$$

Les résidus de Martingale prennent leurs valeurs entre  $-\infty$  et 1 avec des valeurs négatives pour les données censurées.

Ils sont non corrélés dans le cas le cas d'échantillons de tailles importantes.

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Le modèle de Weibull est un modèle à risque proportionnels

Si on désigne par  $\phi$  le facteur d'accélération de la survie dans une population **1** par rapport à celle dans une population de référence **0**, on a la relation suivante entre les fonctions de Survie :

$$S_1(t) = S_0(\phi * t)$$

On en déduit le résultat suivant sur les fonctions de risque instantanés :

$$h_1(t) = \phi * h_0((\phi * t))$$

Le facteur  $\phi$  est une quantité positive. On peut l'exprimer sous la forme d'une exponentielle fonction de la valeur de variables  $X$  codant l'exposition à des facteurs de risque donné :

$$\phi = \exp(X * \beta)$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Le modèle de Weibull est un modèle à risque proportionnels

La fonction de risque instantanée pour une population soumise au profil de facteurs de risques X s'écrit alors :

$$h(t, x) = \exp(x * \beta) * h_0(t * \exp(x * \beta))$$

Où  $h_0(t)$  désigne la fonction de risque de base ( $X = 0$ ) ('**Baseline Hazard Function**')

Si l'on suppose que le temps  $t$  suit dans la population de base une loi de Weibull  $W(\lambda ; \gamma)$

La fonction de risque pour cette population de base s'écrit :

$$h_0(t) = \lambda * \gamma * t^{(\gamma-1)}$$

La fonction de risque pour la population de profil X s'écrit alors :

$$h(t, x) = \exp(x * \beta) * \left[ \lambda * \gamma * (t * \exp(X * \beta))^{\gamma-1} \right]$$

# Séminaire d'Epidémiologie Animale

## Analyses de données de survie (modèles)

---

### Le modèle de Weibull est un modèle à risque proportionnels

La fonction de risque dans la population à risque s'écrit après quelques réarrangements :

$$h(t, x) = (\exp(x * \beta))^{\gamma} * \left[ \lambda * \gamma * (t)^{\gamma-1} \right]$$

Soit : 
$$h(t, x) = (\exp(x * \beta))^{\gamma} * h_0(t)$$

Le rapport des risques est égal à : 
$$HR(x) = (\exp(x * \beta))^{\gamma} = \exp(x * \beta * \gamma)$$

Dans le cas d'un seul facteur de risque binaire ( $X = 1$  (exposé) /  $0$  (non exposé))

Le rapport des risques est égal à : 
$$HR = \exp(\beta * \gamma)$$

Attention !! Il s'interprète comme le rapport des risques du groupe 'non exposé' ( $X=0$ ) par rapport au groupe 'exposé' ( $X=1$ ) (temps de l'événement ( $x=1$ ) < temps ( $x=0$ )  $\rightarrow \beta < 0$ )