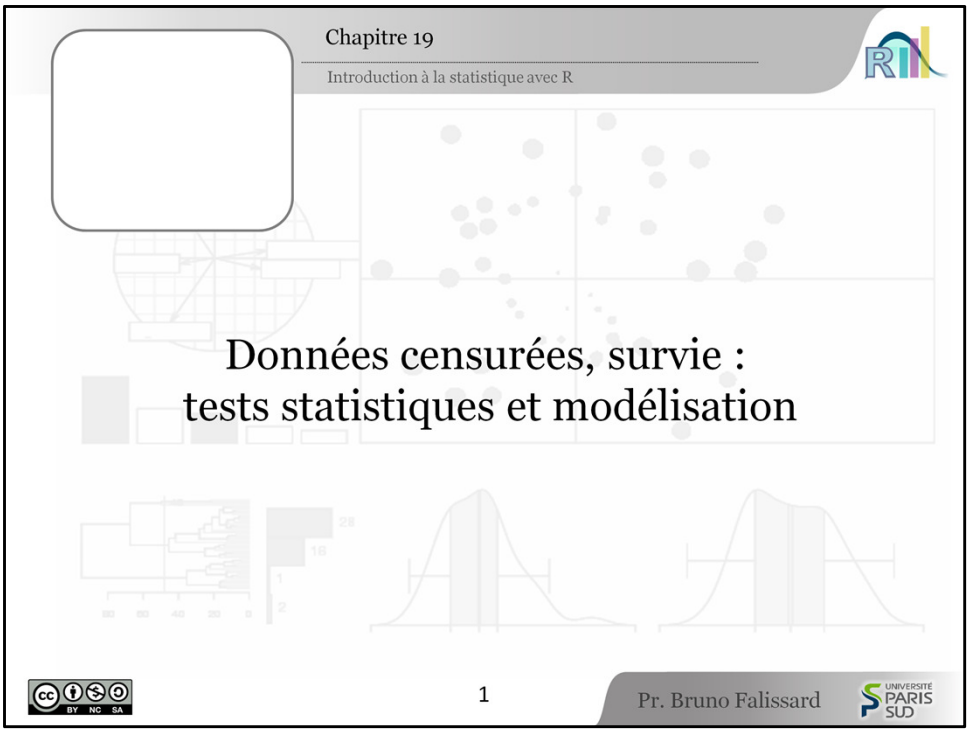
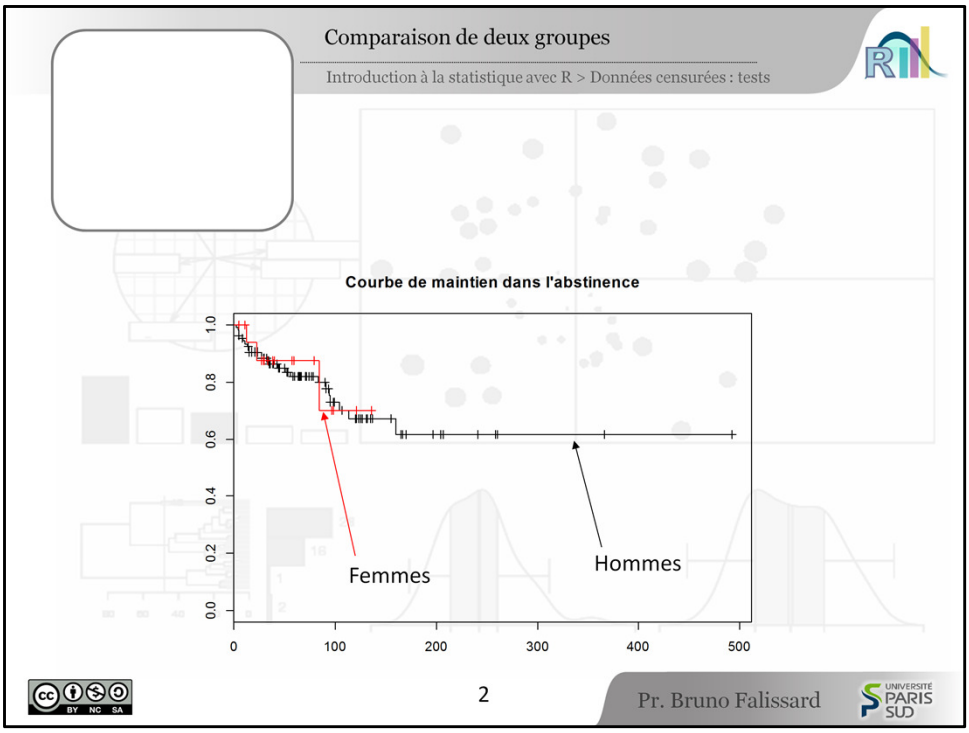
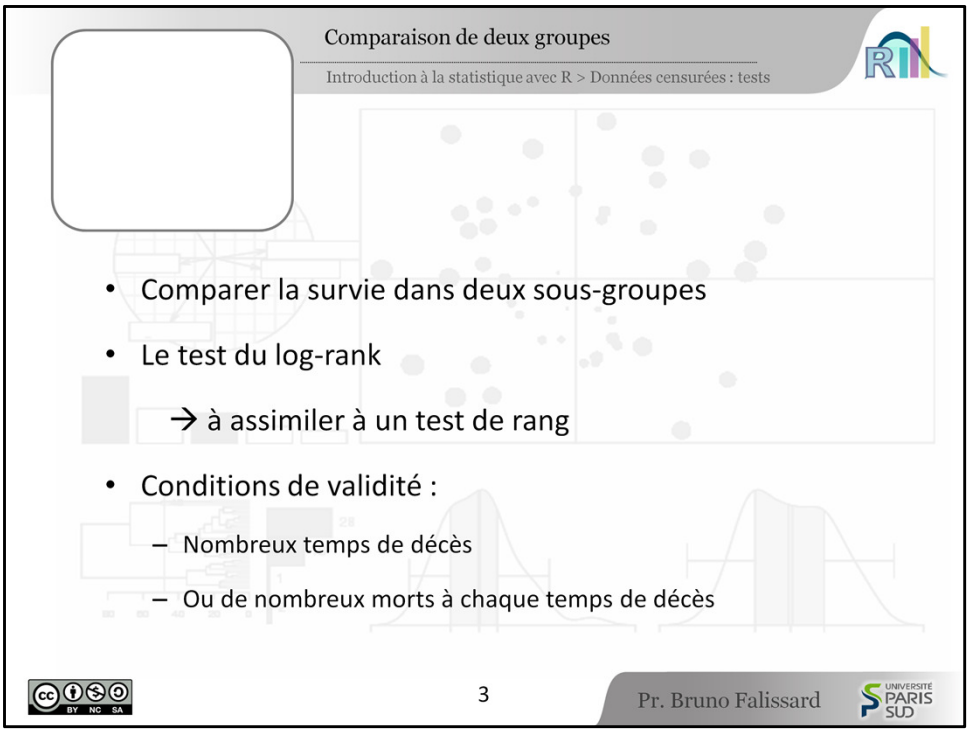


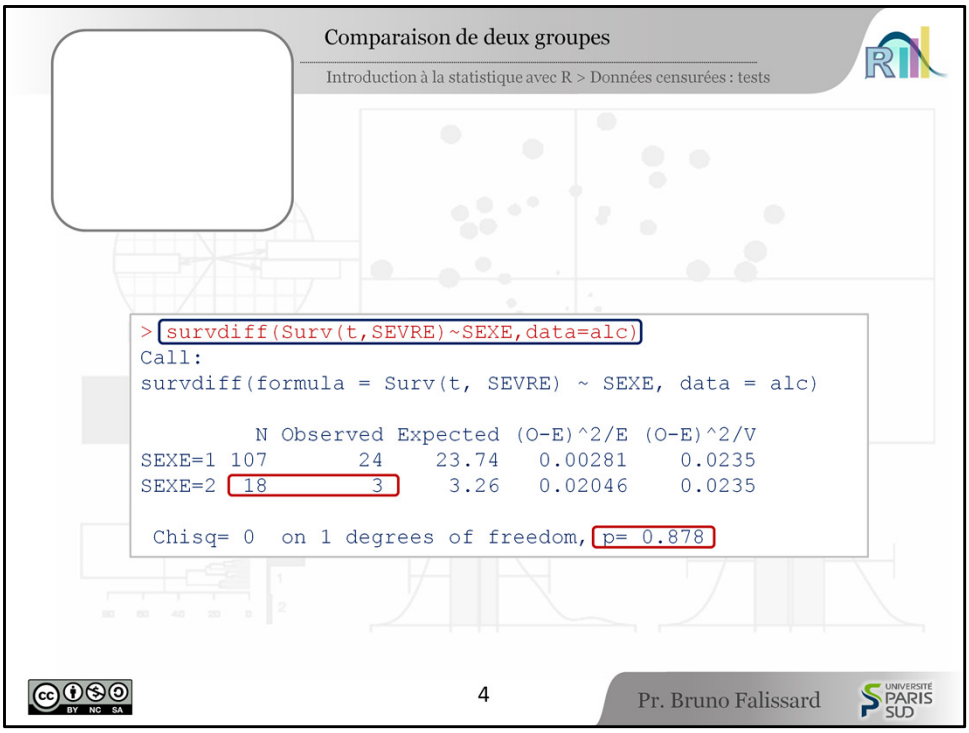
[0:00] Nous allons aborder maintenant le deuxième chapitre sur les données censurées en insistant plus particulièrement sur les tests statistiques et les modèles multivariés.



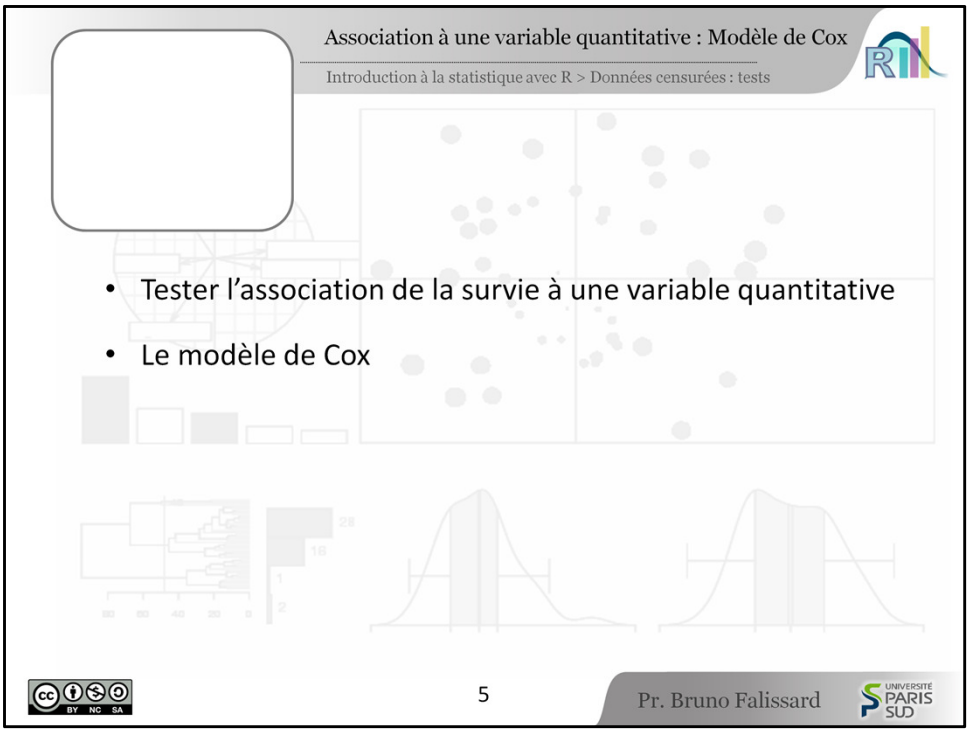
[0:12] A la fin du chapitre précédent, nous avons représenté sur le même graphique, les fonctions de survie relatives à la rechute de la maladie alcoolique chez un groupe de femmes et chez un groupe d'hommes. Les deux courbes avaient l'air superposées. Mais on peut quand même se poser la question et tester statistiquement la différence de taux de rechute entre les hommes et les femmes. Pour réaliser un tel test, il faut utiliser la méthode du Log-Rank.



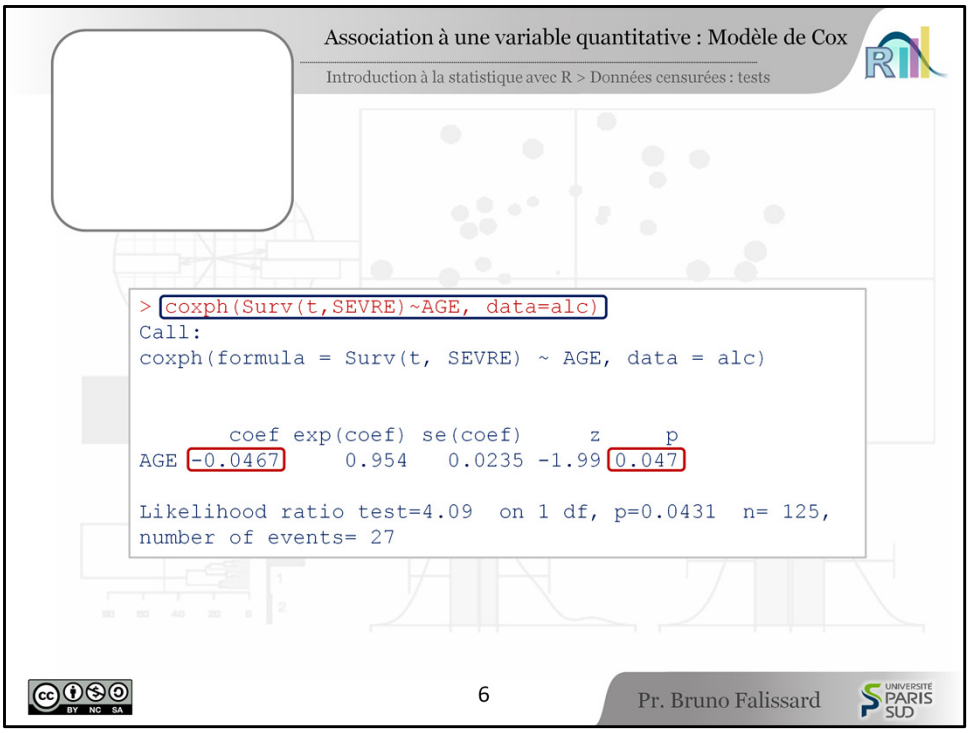
[0:44] Le test du Log-Rank est assimilé à un test de rangs, rangs entre les temps de décès, un peu comme le test de Wilcoxon. Les conditions de validité du Log-Rank sont un petit peu délicates. La principale, c'est s'il y a de nombreux temps de décès alors le test du Log-Rank est valide. Et, c'est à peu près compatible avec ce qui s'est passé avec notre échantillon. Il y a des situations où la question ne se pose pas comme ça. Par exemple, vous suivez des sujets et vous ne les observez que tous les six mois. Alors, vous aurez peu de temps de décès ; mais vous aurez beaucoup de décès à chaque observation, et c'est le deuxième volet des conditions de validité du test du Log-Rank. Quand il y a de nombreux morts à chaque temps de décès, alors il est aussi valide.



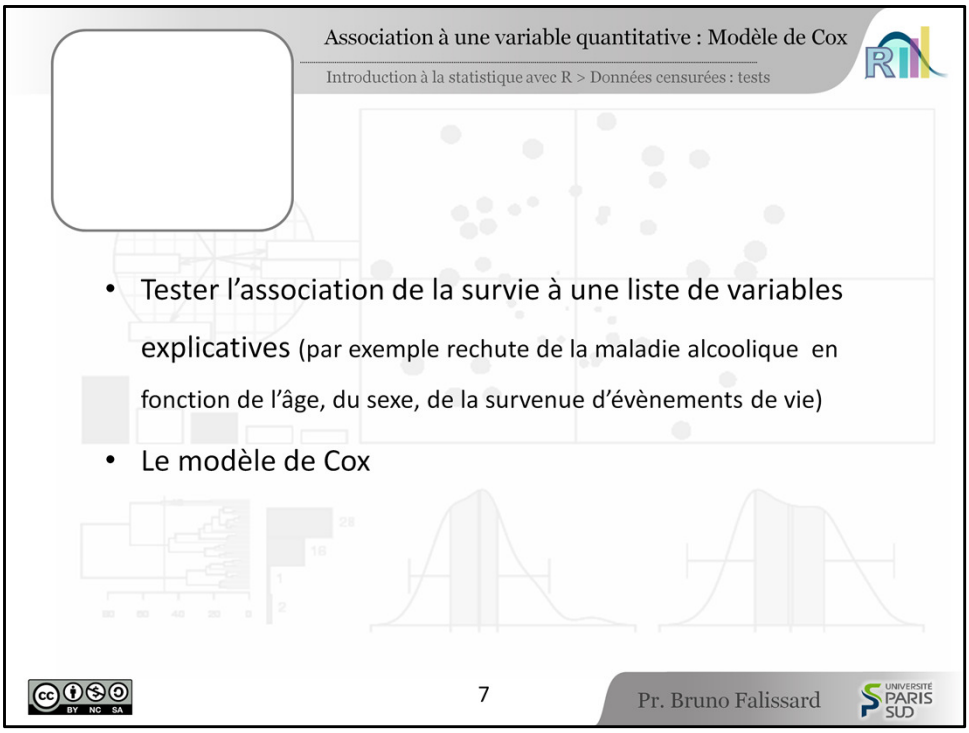
[1:25] Alors, voyons ce que va donner ce test sur notre fichier de données avec R. La syntaxe à utiliser est maintenant assez classique. Il faut utiliser la fonction `survdiff()` et puis on enchaîne avec, comme dans le chapitre précédent, la fonction `Surv()` avec un S majuscule et puis, d'abord la variable délai, puis la variable sevrage . On ferme la parenthèse, tilde, sexe, et puis après on spécifie le fichier de données. Les résultats sont très simples à interpréter. A la dernière ligne, nous avons le `p=0.87`. A l'évidence, il n'y a pas de différence entre le pourcentage de rechute chez les hommes et chez les femmes. Mais, de toute façon, ce test était un peu illusoire. On voit sur la deuxième ligne d'observation à `sexe=2` , `N=18`. Il n'y a que 18 femmes dans notre échantillon et `Observed=3`, c'est-à-dire seulement 3 rechutes chez les femmes. Ces échantillons étaient vraiment trop petits pour avoir une puissance suffisante pour séparer les hommes des femmes.



[2:28] Il y a des situations où il est potentiellement intéressant de tester l'association entre la survie et une variable quantitative. Dans notre exemple, on pourrait tester l'association entre le risque de rechute de la maladie alcoolique et puis l'âge. D'ailleurs, cette association, on pourrait l'imaginer dans les deux sens. D'une part, les sujets jeunes pourraient mal se connaître, sous-estimer le risque de rechute et donc présenter des récidives plus précoces. Au contraire, on pourrait imaginer que les sujets âgés sont des patients chroniques enkystés dans leur maladie alcoolique et donc le risque de rechute pourrait être plus élevé. Dans tous les cas, la méthode statistique qui permet de tester une telle association est le modèle de Cox. Nous verrons dans les diapositives suivantes, comment vérifier ces conditions de validité. Voyons tout de suite, comment estimer un modèle Cox avec R.

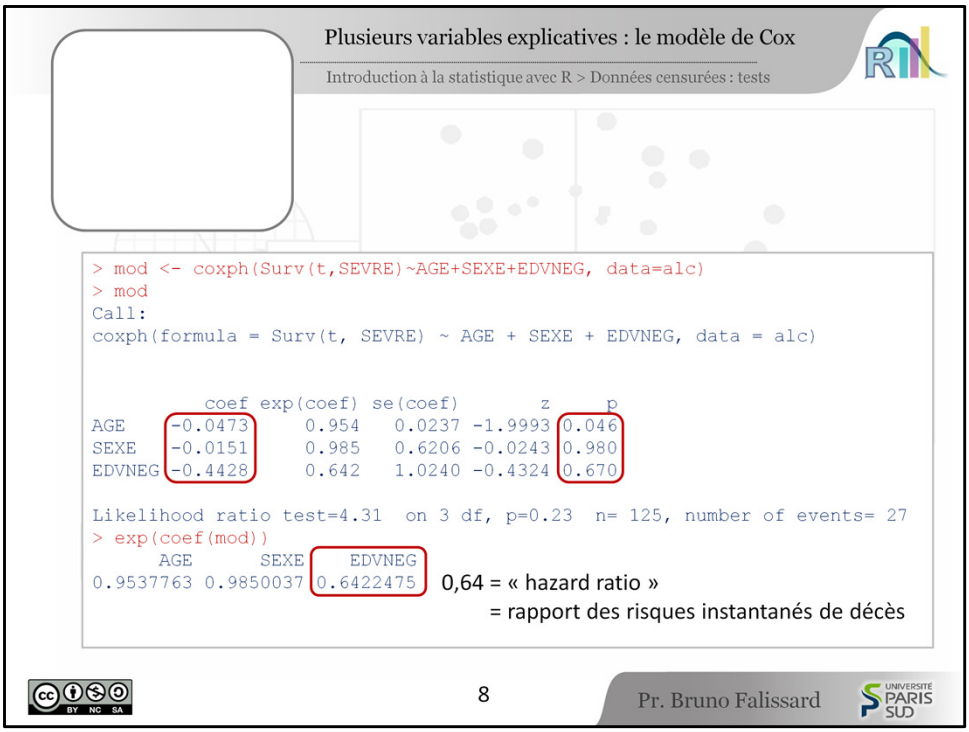


[3:24] Il faut utiliser la fonction `coxph()` et puis on retrouve ensuite, la syntaxe habituelle, `Surv()` avec un grand S, le délai de suivi `t` et la variable rechute `SEVRE`, tilde, `age` et puis le fichier de données `data=alc`. Les résultats sont très faciles à interpréter. Nous avons une seule variable explicative, l'AGE. Au bout de la ligne, on a le `p=0.047`, donc le `p` est tout juste inférieur à `5%` . Mais, au risque de 5%, on peut dire qu'il y a une association significative entre l'âge et le risque de rechute de la maladie alcoolique. Alors, quel est le sens de cette association ? Pour interpréter, pour trouver le sens, il faut aller voir le coefficient. Dans, la colonne `Coef`, nous avons un coefficient qui vaut `-0,0467`. Nous interpréterons plus tard, la taille de ce coefficient. Mais là, nous avons d'abord son signe. Son signe est négatif. Ce qui sous entend que la survenue d'une rechute alcoolique va être plus tardive pour les gens qui sont plus âgés. Donc, l'âge a tendance à protéger de la rechute.



[4:33] Comme dans la régression linéaire multiple ou dans la régression logistique multiple, il est tentant de tester ici l'association entre la survie et une liste de variables explicatives. Avec notre nouveau jeu de données, un exemple naturel serait de tester l'association entre le risque de rechute de la maladie alcoolique et puis, l'âge, le sexe et les événements négatifs pendant le suivi. On aurait ainsi, la force spécifique, la substantifique moelle de chaque variable explicative sur la variable à expliquer.

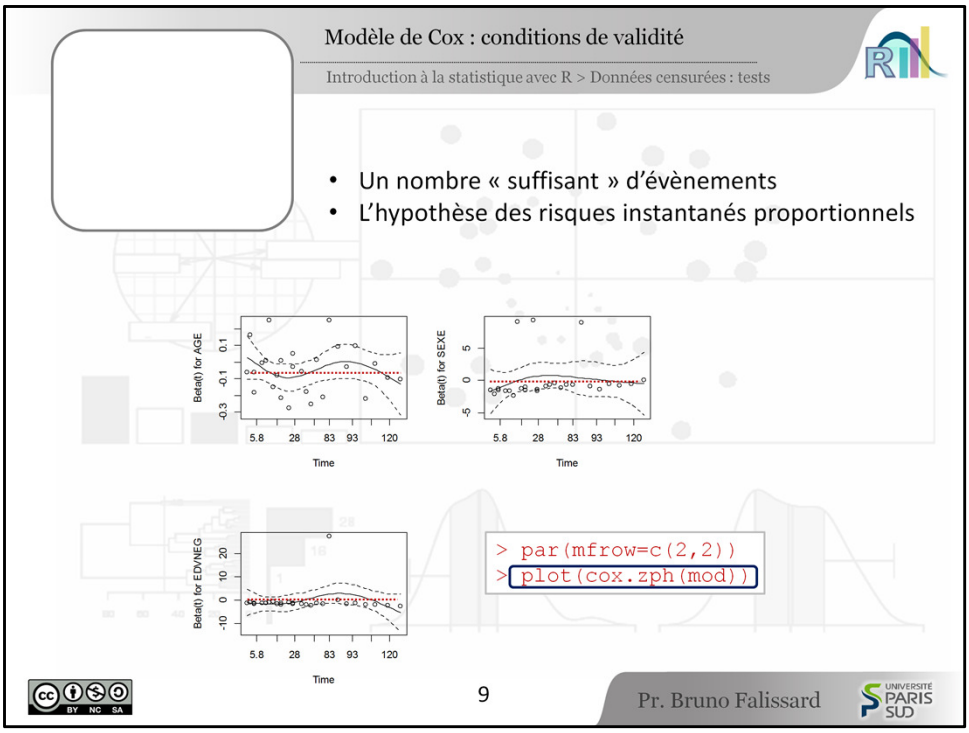
Le modèle qui permet de tester une telle association est aussi le modèle de Cox.



[5:14] La syntaxe à utiliser avec R est très semblable à la syntaxe utilisée dans les dernières diapositives : `coxph(Surv())`, le délai, l'évènement et puis les trois variables explicatives AGE, SEXE, EDVNEG et puis enfin le nom du fichier.

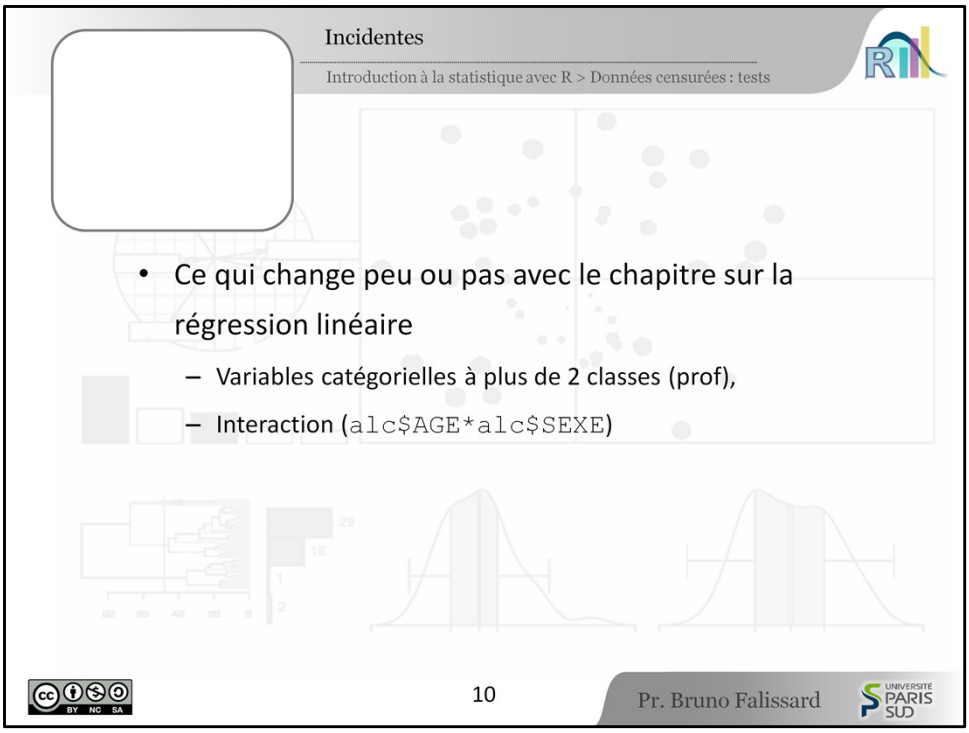
Nous avons maintenant, trois lignes pour les variables explicatives. Au bout des lignes les `p` et nous constatons que seulement l'âge est statistiquement associé au risque de rechute et encore avec une significativité limite. Le sexe et les évènements de vie négatifs ne sont donc pas statistiquement associés au risque de rechute de la maladie alcoolique. Encore que tout ça est à relativiser parce que nous avons vu qu'il y avait très peu de femmes dans cet échantillon, donc très peu de puissance et il y a pour les évènements de vie négatifs, seulement 5 sujets qui en ont eus pendant la durée de suivi. Donc, pour sexe et évènement de vie négatif, la puissance statistique est très faible. Les tests statistiques qui conduisent à accepter l'hypothèse nulle sont donc à interpréter avec beaucoup de prudence. Qu'en est-il de l'interprétation des coefficients ? Ils sont négatifs. Donc, c'est plutôt dans un sens de protection. En tant que tels, ils sont très difficiles à interpréter, en dehors de leur signe. Par contre, l'exponentielle de ces coefficients peut être interprétée, surtout quand la variable explicative correspondante est binaire.

Prenons, l'exemple des événements de vie. L'exponentielle du coefficient vaut `0.64`, nous avons donc, `36%` de chances de moins présenter un risque de rechute à un instant donné. Ce `0.64` correspond à un "hazard ratio" ou "rapport des risques instantanés de décès". Comme nous l'avons dit, ce rapport de risques instantanés qui vaut ici 0. 64, correspond au fait que l'on a 36% (36%, c'est le complément à 100 de 64), 36% de chances en moins de faire une rechute de la maladie alcoolique quand on a eu des évènements de vie plutôt que quand on n'en a pas eu. Bien sûr, ici, ça peut sembler paradoxal. Mais, souvenons-nous qu'il n'y a que 5 sujets qui ont eu des évènements de vie négatifs et que la variable, en tout état de cause, n'est pas statistiquement associée. Donc, l'interprétation de ce "hazard ratio" est purement à titre pédagogique.

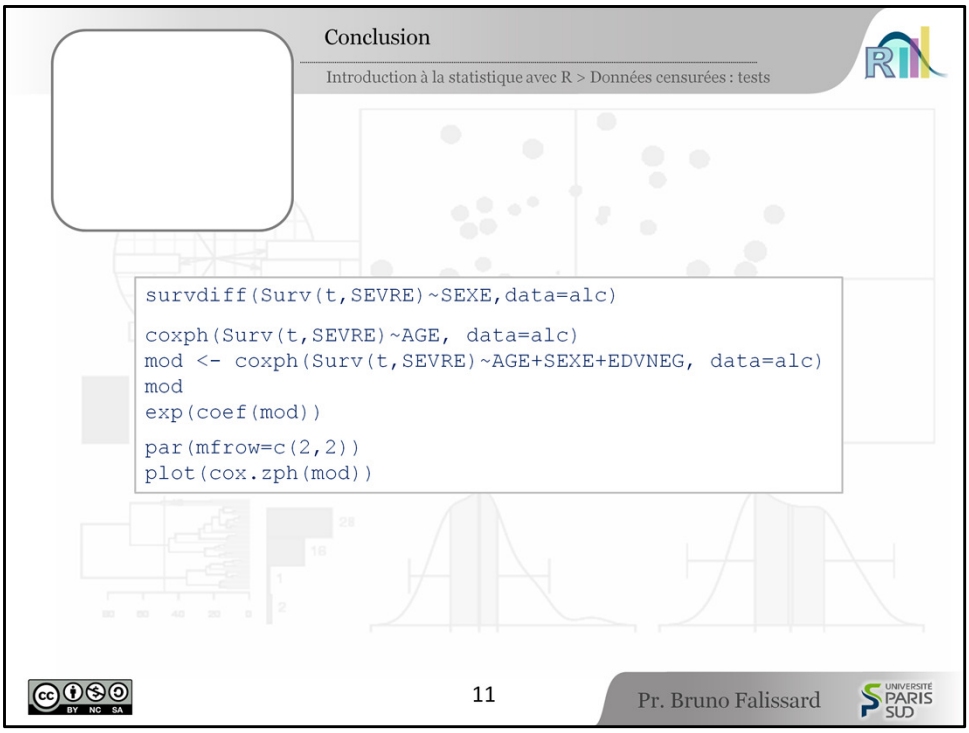


[7:40] Alors, dernier point, capital, incontournable : comment vérifier les conditions de validité du modèle de Cox ? Et là, c'est pas si simple que ça, c'est le moins qu'on puisse dire. Premier élément, peut être le plus facile, comme dans la régression logistique, il faut un nombre suffisant d'évènements, c'est-à-dire 5 à 10 par variable explicative. Je vous renvoie à la diapositive spécifique du cours sur la régression logistique. Alors, maintenant, venons-en à la condition de validité propre au modèle de Cox : vérifier l'hypothèse des risques instantanés proportionnels. Il faut définir à quoi çà correspond. Malheureusement, il faudrait écrire tout un tas d'équations, mais ce n'est pas du tout l'objet du présent cours. Alors, par chance, il est relativement aisé avec R de vérifier graphiquement cette hypothèse des risques instantanés proportionnels. L'instruction est très simple. C'est `plot(cox.zph())` du modèle qui a été estimé. Alors, vous voyez d'abord qu'il y a une instruction `par()`, puis `mfrow()`, quelque chose d'un peu plus compliqué, c'est tout simplement pour fragmenter la fenêtre graphique de R pour pouvoir représenter quatre schémas (2 fois 2). Donc l'instruction `par()` sert à avoir les mêmes schémas d'un seul coup sur la fenêtre graphique de R et puis après nous tapons `plot(cox.zph())` du modèle de Cox qui a été estimé.

Nous obtenons ici, trois graphiques, trois pour les 3 variables : AGE, SEXE et EDVNEG (évènement de vie négatif). Et, nous devons obtenir trois courbes en traits continus qui sont le plus horizontal possible. Ce qui est à peu près globalement le cas ici. Donc, nous dirons, en première approximation, que l'hypothèse des risques instantanés proportionnels est vérifiée. Nous n'insisterons pas dessus.



[9:35] Mais, comme dans le chapitre sur la régression linéaire multiple et sur la régression logistique, on peut inclure dans ces modèles des variables catégorielles à plus de deux classes qui seront recodées automatiquement en variables binaires et mettre des termes d'interaction entre des variables pour rechercher les synergies entre variables explicatives.



[9:49] Test du Log-Rank, modèle de Cox, vérification des conditions de validité du modèle de Cox, voici quelques instructions que je vous invite à refaire sur votre ordinateur, ça vous aidera à mieux comprendre ces quelques notions qui sont quand-même assez délicates.