

I - Le cas de deux populations

On veut ici comparer la survie de deux groupes A et B. Pour ce faire, nous allons comparer les survies de ces deux groupes à un instant t donné, en se rappelant que la survie au temps t correspond à une proportion: on va donc pouvoir approximer la loi binomiale par une loi normale, et ainsi montrer que la statistique

$$\frac{\hat{S}_A(t) - \hat{S}_B(t)}{\sqrt{\widehat{\text{Var}}(\hat{S}_A(t)) + \widehat{\text{Var}}(\hat{S}_B(t))}} \underset{n \rightarrow \infty}{\overset{\mathcal{L}}{\sim}} \mathcal{N}(0,1) \quad \text{sous } H_0: S_A(t) = S_B(t).$$

En effet, sous H_0 , le numérateur est bien centré!

Rp: Par contre, on se rend compte que cette comparaison s'effectue pour un temps t donné, à défaut de permettre de comparer l'«intégralité» des distributions de survie...

→ Dans la suite, on introduit les notations suivantes: basé sur une approche non-paramétrique, le principe des tests à venir est de comparer le nombre de décès observés dans chacun des groupes avec le nombre de décès attendus (obtenus sous H_0):

- $T_1 < \dots < T_N$: les temps de décès ordonnés sur les 2 échantillons réunis.
- d_{Ai} et d_{Bi} : le nb de décès ~~obs~~ observés dans les groupes A et B au temps T_i .
- $d_i = d_{Ai} + d_{Bi}$
- Y_{Ai} et Y_{Bi} : nb de sujets à risque en T_i dans les groupes A et B.
- $Y_i = Y_{Ai} + Y_{Bi}$.

→ On peut donc résumer l'information globale par chaque temps d'événement T_i sous la forme d'un tableau:

en T_i	Nb décès en T_i	Vivants après T_i	expo
Groupe A	d_{A_i}	$Y_{A_i} - d_{A_i}$	Y_{A_i}
Groupe B	d_{B_i}	$Y_{B_i} - d_{B_i}$	Y_{B_i}
	d_i	$Y_i - d_i$	Y_i

II - Statistiques de test avec 2 populations

On rappelle qu'on cherche à tester l'hypothèse nulle selon laquelle les distributions de survie sont les mêmes dans les 2 groupes à un instant t , soit:

$$H_0 : S_A(t) = S_B(t), \text{ pour } t \text{ donné.}$$

→ Ainsi, sous H_0 , la proportion attendue de décès parmi les exposés est la même au temps t . On étend cette égalité pour tous les temps de décès T_i !

→ Ainsi, pour chaque temps T_i , on compare les proportions de décès parmi les sujets à risque dans chacun des groupes à l'aide d'un test du χ^2 ...

① - Formalisation avec variables aléatoires

Soit D_{A_i} (respectivement D_{B_i} et D_i) la v.a. dont la valeur observée est d_{A_i} (respectivement d_{B_i} et d_i). On peut montrer que D_{A_i} suit une loi hypergéométrique.

Rappel : On symbolise une loi hypergéométrique par l'expérience suivante : Prenons N boules dans un sac, et tirons successivement sans remise n boules. Le sac contient $N_1 = pN$ boules gagnantes et $N_2 = (1-p)N$ boules perdantes : $X \sim \text{hypergeo}(N, p)$ est la v.a. comptant le nb de boules gagnantes...
 $\xrightarrow{\text{décès}}$ $\xrightarrow{\text{survivants}}$ D_{A_i}

(2)

→ On remarque que au temps T_i :

$$E[D_{A_i}] = Y_{A_i} \times \frac{d_i}{Y_i} = \text{expo groupe } A_i \times \text{tx de mortalité toutes populations confondues (A et B, sous } H_0).$$

$$\text{Var}(D_{A_i}) = \frac{(Y_i - d_i)}{(Y_i - 1)} \times \frac{d_i Y_{A_i} Y_{B_i}}{Y_i^2} \left[\begin{array}{l} (D_{A_i}^X \sim \text{HyperGéo}(n, p, N)) \\ \downarrow \quad \quad \quad \uparrow \\ d_i \quad \quad \quad Y_i \\ \hookrightarrow E(X) = np \quad \quad \quad \text{Var}(X) = npq \left(\frac{N-n}{N-1} \right) \end{array} \right]$$

→ Interprétation:

• $E[D_{A_i}]$: nb de décès attendus dans le groupe A au temps T_i .

→ Sous H_0 , on montre que $\frac{D_{A_i} - E[D_{A_i}]}{\sqrt{\text{Var}(D_{A_i})}} \underset{n \rightarrow \infty}{\mathcal{L}} \mathcal{N}(0, 1)$, donc $\left(\frac{D_{A_i} - E[D_{A_i}]}{\sqrt{\text{Var}(D_{A_i})}} \right)^2 \sim \chi_1^2$.

→ En considérant des pondérations sur les temps T_i , données par w_i ($i=1, \dots, N$), alors par \perp entre les v.a. D_{A_i} et D_{A_j} (associées à T_i et T_j), on a: (somme sur les T_i)

$$\sum_{i=1}^N w_i (D_{A_i} - E[D_{A_i}]) = \sum_{i=1}^N w_i \left(D_{A_i} - \frac{Y_{A_i} d_i}{Y_i} \right) \underset{n \rightarrow \infty}{\mathcal{L}} \mathcal{N}\left(0, \sum_{i=1}^N w_i^2 \text{Var}(D_{A_i})\right)$$

Ainsi, sous H_0 , les statistiques

$$\chi_0^2 = \frac{\left[\sum_{i=1}^N w_i \left(D_{A_i} - \frac{Y_{A_i} d_i}{Y_i} \right) \right]^2}{\sum_{i=1}^N w_i^2 \frac{(Y_i - d_i)}{(Y_i - 1)} \frac{d_i Y_{A_i} Y_{B_i}}{Y_i^2}} \underset{n \rightarrow \infty}{\mathcal{L}} \left(\mathcal{N}(0, 1) \right)^2 \underset{n \rightarrow \infty}{\mathcal{L}} \chi_1^2$$

② - Le test du log-rank:

Sans doute le test le plus utilisé et le plus connu.

Il consiste à reprendre la statistique ci-dessus en fixant $w_i = 1$.

→ Interprétation: on attribue à chaque décès la même importance, quelle soit l'instant où il survient. Le test compare pr chaque décès le nb de décès observés à celui-attendus.

③ - Le Test de Gehan:

Un peu moins connu, mais tout de même largement utilisé.

Il consiste à fixer $w_i = Y_i$ dans la statistique. Cela signifie que les termes prenant le plus d'importance dans le test sont ceux pour lesquels Y_i est grand.

⇒ Les poids sont donc plus élevés pour les décès précoces que les décès tardifs!

④ - Le Test de Peto et Prentice:

Le moins connu. Dans cette version du test, on fixe $w_i = \frac{\frac{1}{n} \sum_{k=1}^i Y_k}{Y_k + d_k}$

En réalité, ces pondérations sont proches de l'estimateur de Kaplan-Meier (de la fonction de survie). Comme pour le test de Gehan, on accorde donc plus de poids aux décès précoces, puisque la fonction de survie est d'autant plus élevée que i est petit.

→ Rq général: • Il existe une version approchée du test du log-rank pour un calcul "à la main".

• Par construction, ces tests ne fonctionnent que si les courbes de survie des deux groupes ne se croisent pas sur l'ensemble de la période étudiée! (sinon la puissance du test ↓).

III - Cas avec plus de 2 populations.

On peut aisément étendre l'ensemble des tests vu précédemment au cas de K populations à étudier, où $K > 2$.

Nous ne verrons ici que l'extension du test du log-rank, car c'est le plus utilisé.

Pour simplifier la présentation, nous prendons le cas où nous disposons de 3 échantillons, i.e. $K=3$.

Ainsi, on peut résumer par chaque temps de décès T_i l'information dans le tableau suivant:

	Décès en T_i	Vivant après T_i	Expo
Groupe A	d_{A_i}	$Y_{A_i} - d_{A_i}$	Y_{A_i}
Groupe B	d_{B_i}	$Y_{B_i} - d_{B_i}$	Y_{B_i}
Groupe C	d_{C_i}	$Y_{C_i} - d_{C_i}$	Y_{C_i}
	d_i	$Y_i - d_i$	Y_i

En utilisant exactement le même raisonnement que précédemment avec deux échantillons, on peut montrer que le vecteur $V = \begin{pmatrix} \sum_{i=1}^N D_{A_i} - E[D_{A_i}] \\ \sum_{i=1}^N D_{B_i} - E[D_{B_i}] \end{pmatrix}$ est un vecteur gaussien en dimension 2, avec :

$$\mu = \begin{cases} E[D_{A_i}] = \frac{Y_{A_i} d_i}{Y_i} \\ E[D_{B_i}] = \frac{Y_{B_i} d_i}{Y_i} \end{cases} \quad \text{et} \quad \Sigma = \begin{cases} \text{Var}(D_{A_i}) = \frac{Y_i - d_i}{Y_i - 1} \frac{d_i Y_{A_i} (Y_i - Y_{A_i})}{Y_i^2} \\ \text{Var}(D_{B_i}) = \frac{Y_i - d_i}{Y_i - 1} \frac{d_i Y_{B_i} (Y_i - Y_{B_i})}{Y_i^2} \\ \text{Cov}(D_{A_i}, D_{B_i}) = -\left(\frac{Y_i - d_i}{Y_i - 1}\right) \frac{d_i Y_{A_i} Y_{B_i}}{Y_i^2} \end{cases}$$

$$\text{Ainsi } V_2^* = \frac{V}{\sigma} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_2^* \right) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} -$$

On en déduit que la statistique $\chi^2_{\text{sous } H_0} = V^T \begin{pmatrix} \sum_{i=1}^N \text{Var}(D_{A_i}) & \sum_{i=1}^N \text{Cov}(D_{A_i}, D_{B_i}) \\ \sum_{i=1}^N \text{Cov}(D_{A_i}, D_{B_i}) & \sum_{i=1}^N \text{Var}(D_{B_i}) \end{pmatrix}^{-1} V$

suit une loi du Chi-2 à 2 degrés de liberté! (car contient une somme de 2 termes gaussiens centrés réduits au carré).

→ Après avoir mené ce raisonnement avec le couple (A,B), on peut le décliner avec les couples (A,C) et (B,C).

⇒ On obtient d'autres statistiques de test équivalentes!

Généralisation:

Lorsque l'on étudie la survie de k groupes, et que l'on désire tester:

$[H_0]: S_1(t) = S_2(t) = \dots = S_k(t)$ (vs) $[H_1]:$ au moins 2 fonctions de survie différentes,

la statistique de test s'obtient de la même façon, et:

$$\chi^2_{n \rightarrow \infty} \sim \chi^2_{k-1} \quad (\text{soit une loi du chi-2 à } (k-1) \text{ degrés de liberté}).$$

• Remarque: Si l'on sait par exemple que la différence entre deux survies peut s'expliquer par un facteur de confusion (cf plus tard dans le cours), on peut comparer les fonctions de survie en ajustant sur ce facteur et en utilisant la même procédure. Cela consiste à comparer les fonctions de survie pour une valeur fixée (donnée) de ce facteur de confusion.

⇒ Ce test est souvent dénommé le "test du logrank stratifié".

(Chaque strate constitue un sous-échantillon à partir duquel il est possible de calculer une différence entre le nb de décès observés et attendus. On peut ensuite faire la somme sur chacune des strates (les individus étant \neq entre strates, les v.a. sont \perp)).

→ Illustrations R!